Universidad del País Vasco

Euskal Herriko Unibertsitatea

Universidad del País Vasco / Euskal Herriko unibertsitatea

Doctoral Thesis

Implementation of computational techniques to study protein synthesis and folding.

Author: Óscar Rodríguez Ballesteros
Supervisor: Aritz Leonardo Liceranzu
Co-supervisor & Tutor: Aitor Bergara Jauregui

January 2023

Acknowledgements

I want to thankfully acknowledge the Materials Physics Center for providing me with financial support and giving me the opportunity to work in this thesis and continue my academic career, it has been a pleasure to be part of the research center and I want to thank all the administrative team as they have helped me through all bureaucracy easing a lot my stay here. In a similar manner, I want to express my gratitude towards the Donostia International Physics Center (DIPC), since they led me use ATLAS-EDR and ATLAS-FDR supercomputers throughout these four years, as well as providing me with technical support when needed. Also, I want to acknoledge the Spanish network of high-performance computing centers (Red Española de Supercomputación, RES), since we made use of their computational resources, in particular, those at the Turgalium center under the project (RES-BCV-2022-0002) with the highly appreciated technical support provided by CETA-CIEMAT.

Después de los agradecimientos formales, me gustaría ya agradecer a mis tutores, Aitor y Aritz, y a Rafael el haberme guiado a lo largo de estos años, ha sido bonito ver como hemos empezado y como hemos acabado. También tenemos que ser sinceros y decir que nuestra incursión en el mundo de la biología no habría sido posible sin la paciencia de Álvaro y Arantza, así como del resto de miembros del laboratorio kcnq del instituto Biofisika.

List of Publications

- Janire Urrutia et al. "An epilepsy-causing mutation leads to co-translational misfolding of the Kv7.2 channel". In: BMC biology 19.1 (May 2021). PMC8138981[pmcid], pp. 109–109. issn: 1741-7007.

  doi: 10.1186/s12915-021-01040-1.

- Arantza Muguruza-Montero et al. "Do calmodulin binding IQ motifs have built-in capping domains?" In: Protein Science 30.10 (2021), pp. 2029–2041.

  doi: https://doi.org/10.1002/pro.4170.

- Rafael Ramis et al. "Molecular dynamics simulations of the calmodulin-induced α-helix in the SK2 calcium-gated potassium ion channel" In: Journal of Biological Chemistry, Volume 299, Issue 2, 102850.

  doi: https://doi.org/10.1016/j.jbc.2022.102850

- Improved KCNQ2 gene missense variant interpretation with artificial intelligence (preprint article).

Contents

List of Figures

Abstract

Abstract in English

Throughout this thesis, two different biological problems have been studied, chapters 2, 3 and 4 are dedicated to the analysis of the cotranslational folding of the ion channels $K_V7.2$ and SK2, whereas chapter 5 is a study of the allosteric communication within these ion channels and the effect of some channel activators on these allosteric signaling.

The introduction to cotranslational folding starts with the analysis of the W344R mutation of the $K_V7.2$ channel, for which recent experimental assays suggest that this mutation has an effect on the cotranslational folding of the channel [1]. A computational analysis of this variant of the channel reveals two different conformations for the mutated arginine that could explain the different behaviors of the variant observed computationally. However, recent experimental evidence suggests that calmodulin (CaM) plays a fundamental role in the cotranslational folding. Therefore, chapter 3 focuses on the study of how some CaM targets exhibit some helical content when immersed in water, where this alpha-helical content could subsequently be recognized by CaM; the $K_V7.2$ and SK2 channels are included among those targets analyzed. The induction of additional alpha-helical structure on the SK2 channel in the presence of CaM is also described. In chapter 4, we revisit the $K_V7.2$ channel to apply the methodology introduced in chapter 3 to the W344R mutation, and to study the secondary structure formation of this variant both when immersed in water and cotranslationally.

Finally, in chapter 5, graph theory is applied to analyze molecular dynamics simulations (MD) of the whole $K_V7.2$ and SK4 channels structures, in the presence and in the absence of some channel activators, retigabine and ZTZ240 in the case of $K_V7.2$ and riluzole in the case of SK4, with the aim of determining how these drugs affect the allosterical signaling with the pore. Additionally, a thorough description of the

riluzole's binding pocket is presented, since no experimental structure was available and the pocket had to be modeled.

Resumen en Castellano

A lo largo de esta tesis se han estudiado dos problemas biológicos diferentes, los capítulos 2, 3 y 4 se han dedicado al plegamiento cotraduccional de los canales $K_V7.2$ y SK2, mientras que el capítulo 5 se centra en describir las señales alostéricas que tienen lugar en ambos canales y en particular cómo se pueden ver afectadas por algunos activadores de estos canales.

El estudio del plegamiento cotraduccional se vió motivado por una mutación en particular del canal $K_V7.2$, la W344R, para la que estudios recientes experimentales sugieren que tiene un efecto en el plegamiento cotraduccional, pero que parece no afectar al canal en experimentos *in vitro*. Gracias al análisis computacional mediante dinámica molecular realizado en el capítulo 2, se encontraron dos conformaciones diferentes para la arginina mutante, las cuales podrían explicar los diferentes comportamientos observados experimentalmente. Sin embargo, recientes experimentos sugieren que la calmodulina (CaM) juega un papel fundamental en el plegamiento cotraduccional. Por tanto, en el capítulo 3 se estudia mediante *Hamiltonian replica exchange* como algunas de las dianas de la calmodulina, los canales iónicos SK2 y $K_V7.2$ entre otras, muestran contenido helicoidal cuando están inmersas en agua, donde esta estructura secundaria podría ser posteriormente reconocida por la CaM. También se describe como la calmodulina induce la formación de alfa-hélice en el canal SK2. En el capítulo 4 se vuelve al canal $K_V7.2$ para aplicar esta metodología al estudio de la mutación W344R con el fin de ver su efecto en la estructura secundaria inmersa en agua y de forma cotraduccional.

Finalmente, en el capítulo 5, se aplica la teoría de grafos a simulaciones de dinámica molecular (MD) de los canales completos $K_V7.2$ y SK4 en presencia y ausencia de varios activadores, retigabina y ZTZ240 en el caso de $K_V7.2$ y riluzole en el caso de SK4, con el fin de ver cómo estos fármacos afectan a la señalización alostérica con el poro. También se lleva a cabo unadescripción más exhaustiva del sitio de unión del riluzole dado que se tuvo que modelizar al no haber una estructura resuelta con el fármaco.

Chapter 1

General introduction

This project was motivated by the study of the cotranslational effect of the W344R mutation in the $K_V7.2$ channel (a detailed description of the channel and this mutation is presented in chapter 2) by using *in silico* or computational approaches. Given the solid state physics background of my tutors, we were not aware of the magnitude and the complexity of the problem, but we have tried to make significant advances in this matter together with side projects that came in. Before starting the scientific introduction the structure of this thesis will be briefly summarized and some information will be provided to help the reader to assimilate all the information presented herein.

This thesis is structured in chapters; each of them refers to a different project or group of results, aside of this one, which is a short general introduction spanning molecular dynamics simulations (MD) and ion channels from a biological perspective. Due to the large amount of simulations, supplementary information is given in a github repository accessible through clickable links[1], so it is recommendable that this thesis is read in some device with internet connection. For the sake of reproducibility and in case that someone might be interested, input files for each step of the workflow of all simulations are provided in the online repository, we recommend the reader to look for the files called Supp_info.pdf as they contain the information about the arrangement of the files of each simulation. Also, a guide of the supporting information is provided as an appendix, in which all the files provided are listed for all different simulations performed in this thesis.

[1] All links in the text will be colored in blue.

The main topic of this thesis is the cotranslational folding of ion channels, in particular, $K_V7.2$ and SK2. In chapter 2, we will study the mutation W344R and we will present the results that were published in [1]. Furthermore, we will provide other calculations done afterwards that have not been published but we think that thay might be illustrative and reinforce our hypothesis, this is that the mutated arginine in the mutation W344R is able of alternate between two states whereas the native tryptophan is not. This continuation of the results published in [1] even allowed us to define some experiments to further test this hypothesis, but did not succeeded, making us think that maybe this hypothesis was somehow oversimplifying the problem.

Other experimental results presented in [1] indicated that calmodulin (CaM) had also a cotranslational effect. Therefore, in chapter 3, we wanted to include the CaM into our hypothesis, even though the CaM recognition process is extremely complex. However, in the case of SK2, a nuclear magnetic resonance (NMR) structure was published in which the binding domain of the channel exhibited a helical turn when immersed in water [2], making the authors think that this helical turn would be subsequently recognized by CaM. We used this hypothesis as a framework for studying the appearance of helical turns in water and the subsequent helix induction in the SK2 channel.

Afterwards, in chapter 4 we apply the same framework of chapter 3 to the W344R mutation , and we evaluated the effects provoked by this mutation when the system is immersed in water and in the ribosomal tunnel.

Finally, the last chapter is a side project motivated by a recent new technique to analyze allosteric signaling within proteins, called network analysis, in which graph theory is applied to compute the centrality of the different residues, providing with a magnitude related to the importance of each amino acid in the protein for allosteric signaling. In this chapter, full channel structures of $K_V7.2$ and SK4 were simulated, both in the presence and the absence of some activators of the channels, namely Retigabine and ZTZ240 for $K_V7.2$ and Riluzole for SK4. The aim of this simulations was to analyze the signaling pathways of the channels leading to the opening of the pore and to see how the drugs affect them. In the case of SK4, a thorough description of the binding pocket of Riluzole is provided, since the full channel structure with the drug is not experimentally resolved and we had to place the drug by means of computational tools, and the analysis of the pocket revealed a possible mechanism for the effect of the drug, that is also in agreement with some experimental results [3].

After this small review of the main topics, let us start this thesis with an introduc-

tion of the *in silico* or computational simulations of biological systems, in particular molecular dynamics (MD), followed by a small introduction to ion channels.

1.1 *In silico* calculations of biological systems

A simulation or an *in silico* calculation can be observed as a computer experiment, consisting in the application of a theoretical model to a target system by an algorithm, being the latter a set of static instructions. The action of the theory through the algorithm to the target system will arise an outcome that should be analyzed similarly to an experiment, which means that some validation will be required, as the theory that we introduce in the simulations will always be built to be as accurate as possible but very aware of our technical limitations, like for example the computational resources available. In the case of biology, typically, simulations will complement the information extracted from other experiments, and there it will display all its potential as they have a very different approach to the events under study, based on the movements and interactions of the atoms constituting the system, whereas experiments usually are indirect, meaning that they collect evidences or side effects related to the events of interest but sometimes missing the atomic mechanism.

Generalizing, two main different frameworks can be stablished for biological computing simulations. When simulations are done together with experiments they can be used to validate them by seeing part of the target biological process or by computing the same quantities as in the experiments, elucidating the role of each part of the system with atomic resolution, for example molecular dynamics simulations (MD) were able to observe some mechanisms and intermediate states of ionic channels [4, 5] . Conversely, if the simulations are done before experiments, they can be used as a first filter to direct the experiments and discard those that are less likely to succeed, one example is drug development, where docking simulations helped to identify new chemotypes relevant to opioid analgesic discoveries [6] or massively tested small ligands to an influenza receptor [7].

Frequently, the starting point of simulations is the structural information of the systems of interest, but for the calculations done in this work, atomic precision is needed and this is very challenging from an experimental perspective. Atomic structural determination of proteins was led by nuclear magnetic resonance (NMR), but this approach is limited by the fact that the target molecule needs to be crystallized, which can be troublesome [8], although the field is significantly growing [9]. In the

last few years, structural determination is experiencing a revolution by recent advances in Cryo Electron Microscopy (Cryo-EM) [10]. In a nutshell, one aqueous sample is frozen extremely quickly and its molecules adopt a vitreous phase that does not disrupt macromolecular structures, then, many images of the molecule are collected from different angles to reconstruct the three-dimensional structure of the molecule, with this approach inconceivable large structures have been resolved to an outstanding resolution, like the ebola virus [11], eukaryotic ribosomes [12], or membrane proteins like ion channels [13]. Cryo-EM has the disadvantage that only information of stable domains can be obtained, since the molecule necessarily has to adopt the same conformation in most of the collected images in order to obtain a clear image of its three-dimensional shape, this makes theses structures to only contain static information, leaving the door open to complementary calculations capable of observe the dynamical processes experienced by the biological entities, as it is the case of molecular dynamics (MD) simulations [14]. In this introduction, we will focus on MD simulations, since they comprise most of the calculations done for this thesis.

The last ingredient that has helped simulations, and in particular MD, to experience a mayor gain in popularity and visibility is the increasing computational power accessible to research groups. Putting aside the development of the so-called high-performance computer facilities, which is outstanding, it is noteworthy the evolution experienced by the two main MD software used in this thesis, NAMD [15] and GROMACS [16]; both have made enormous efforts to optimize the resources of computers by allowing parallelization and the use of graphic processing units (GPUs), so that the load of the calculations is balanced. These advances have been taking place throughout the time period of this thesis and have helped to democratized the technique as interesting simulations can be made by almost any research group with a workstation, in particular, our group has seen how our computational power has increased several times.

1.1.1 Molecular dynamics

So, what is MD? In the following we will introduce the physics behind a molecular dynamics simulation of a biological system, trying to understand in the process the approximations used and the information that can be collected through a simulation. Essentially, MD simulation computes the time evolution of a system; to achieve that we compute Newton equation for each particle i:

$$\mathbf{F}_i = -\nabla V = m_i \ddot{\mathbf{r}}_i \tag{1.1}$$

In practice, this equation cannot be integrated (solved) as the number of particles in the system is usually enormous and computing the exact potential experienced by the i-th particle is not trivial at all, we will discuss later the shape of this potential, now we will focus on how to efficiently approximate the solutions of Eq. (1.1). For that purpose we will assume that the potential only depends on the position of the particles constituting our system, since no time dependent perturbations will be introduced. So, for computing the time evolution of the system we can make use of a variant of the most common integration algorithm, which is the velocity Verlet algorithm [17]. The main idea is to make a Taylor expansion for the positions at time $(t + \Delta t)$:

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \dot{\mathbf{r}}(t)\Delta t + 1/2\ddot{\mathbf{r}}(t)\Delta t^2 + O\Delta t^3 \tag{1.2}$$

that provides us with the positions at the following time step $\mathbf{r}(t + \Delta t)$ with an error proportional to Δt^3, as positions, velocities and accelerations at time t are known. These new computed positions are used to define the potential at the following step, $V(\mathbf{r}(t + \Delta t))$, since we have assumed that the potential only depends on the positions of the particles of the system, notice that the error in the potential results from the propagation of the error of the positions; the acceleration will be, from the application of Eq. (1.1):

$$\mathbf{a}(t + \Delta t) = -1/m\nabla V(\mathbf{r}(t + \Delta t)) + O\Delta t^3 \tag{1.3}$$

For clarity, before writing the final expression for the velocities, let's see the Taylor expansions of $\mathbf{v}(t + \Delta t)$ and $\mathbf{a}(t + \Delta t)$:

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \mathbf{a}(t)\Delta t + \frac{1}{2}\dot{\mathbf{a}}\Delta t^2 + O\Delta t^3 \tag{1.4}$$

$$\mathbf{a}(t + \Delta t) = \mathbf{a}(t) + \dot{\mathbf{a}}\Delta t + O\Delta t^2 \tag{1.5}$$

Separating $\dot{\mathbf{a}}$ in Eq. (1.5) and inserting in into Eq. (1.4), we get the expression for the velocities with an error of $O\Delta t^3$:

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t) + \frac{\mathbf{a}(t) + \mathbf{a}(t + \Delta t)}{2}\Delta t + O\Delta t^3 \tag{1.6}$$

Where positions and velocities at step t are assumed to be known. At $t = 0$ starting velocities need to be assigned, usually from a Boltzmann distribution for a

desired temperature. In this derivation of the velocity-Verlet algorithm lies the essence of a MD simulation: velocities and positions are known at time t, therefore the position dependent potential is also known. With Eq. (1.2) the positions at the following time step $t + \Delta t$ are computed, these new positions define the new potential of the next time step, which is used to define the new accelerations with Eq. (1.3), and the new velocities with Eq. (1.6). Finally, new velocities and accelerations are used to compute the coordinates at $t + 2\Delta t$ again with Eq. (1.2). This is repeated until the simulation is completed at time $t + N\Delta t$, providing us with a trajectory of the system with a time resolution of Δt. Notice that this simple and fast algorithm has an error associated with the truncation of the Taylor expansion that goes with Δt^3, where the time step is usually small.

1.1.2 Dealing with the potential

In the previous section we have seen that we have a very powerful algorithm to integrate the equations of motion and generate a trajectory of the system, one crucial question that we left open is how to compute the potential of our system. First of all, we have to determine the resolution of the model, which will establish the accuracy of the outcome of the simulation. For example, if we are modeling a chemical reaction where we want to reproduce the transfer of one electron between two molecules, we will need a quantum model to consider electrons, one of the most popular for that matter is density functional theory [18], but the complexity of the calculations and the addition of the active electrons to the total number of interacting particles substantially increases the computational cost for computing the potential of the system. In fact, in this thesis we will be dealing with biological systems immersed in water, with numbers of atoms ranging from the thousands to the hundred thousands, making this approach impractical. Furthermore, to assure numerical stability, the time step for the time evolution has to be 1-2 femtoseconds, in order to be 5-10 times smaller than the vibration frequency of the C-H bond, the highest one that can be found in a biological system that explicitly includes all the atoms of the systems (all-atom MD).

Therefore, we have to develop an efficient way to compute potentials that fits our interests and lies in the sweet spot between accuracy and computational cost, this is usually accomplished by increasing the specificity of the model. As mentioned before, we will work with biological systems and we will exploit their characteristics. One key feature is that the processes do not include chemical reactions or the breaking

of covalent bonds, therefore a quantum level description of the systems will not be needed as electrons will not play any explicit role. Besides, the systems that we will consider in this thesis (i.e. proteins and cell membranes), are made out of a rather small number of building blocks, in the case of proteins these will be the 20 essential amino acids and for membranes we will use just one lipid, this means that there is a large (but finite!) number of different interactions that can take place in our system, so one intelligent approach is to parametrize all of them and put all these parameters in a list, also known as force field, leaving the prediction ability of the model to the quality of this set of parameters, which are obtained from quantum mechanical simulations and fitted to reproduce experimental properties of proteins or lipids. Some of the most popular ones are AMBER [19], OPLS [20] or CHARMM [21], being the latter the one used in all simulations of this thesis.

Force fields

Force fields are the mathematical tool used to compute the potential energy for biological systems, they consist of the combination of an additive energy function and a set of parameters. Most force fields share a common energy function so that they are referred to as Class I force fields [22]:

$$V(\mathbf{r}_1...\mathbf{r}_n) = V_B(\mathbf{r}_1...\mathbf{r}_n) + V_{NB}(\mathbf{r}_1...\mathbf{r}_n) \tag{1.7}$$

Where the bonded term $V_B(\mathbf{r}_1...\mathbf{r}_n)$ is the sum of some simple geometrical terms:

$$\sum_{bonds} \frac{1}{2}K_b(b-b_0)^2 + \sum_{angles} \frac{1}{2}K_\theta(\theta-\theta_0)^2 + \sum_{imp\,dih} \frac{1}{2}K_\psi(\psi-\psi_0)^2 + \sum_{dihedrals} \frac{1}{2}K_\phi(1+cos(n\phi-\delta)) \tag{1.8}$$

These terms are depicted in Fig. (1.1) and refer to the bond length b, angle θ, torsion angle or improper dihedral ϕ, and out of plane distortion or dihedral ψ, finally $K_b, b_0, K_\theta, \theta_0, K_\psi, \psi_0, K_\phi, n$ and δ are parameters stored in the force field. This set of parameters takes into account the specificity of the interacting atoms, considering not only the different nature of those, that is, if the particle is a carbon or an oxygen, but also the environment surrounding them, for instance differentiating carbon atoms belonging to the protein backbone from others in the side chains.

The non bonded term $V_{NB}(\mathbf{r}_1...\mathbf{r}_n)$ is the sum of a Lennard-Jones type potential that accounts for the Van der Waals interactions and a Coulomb term for the

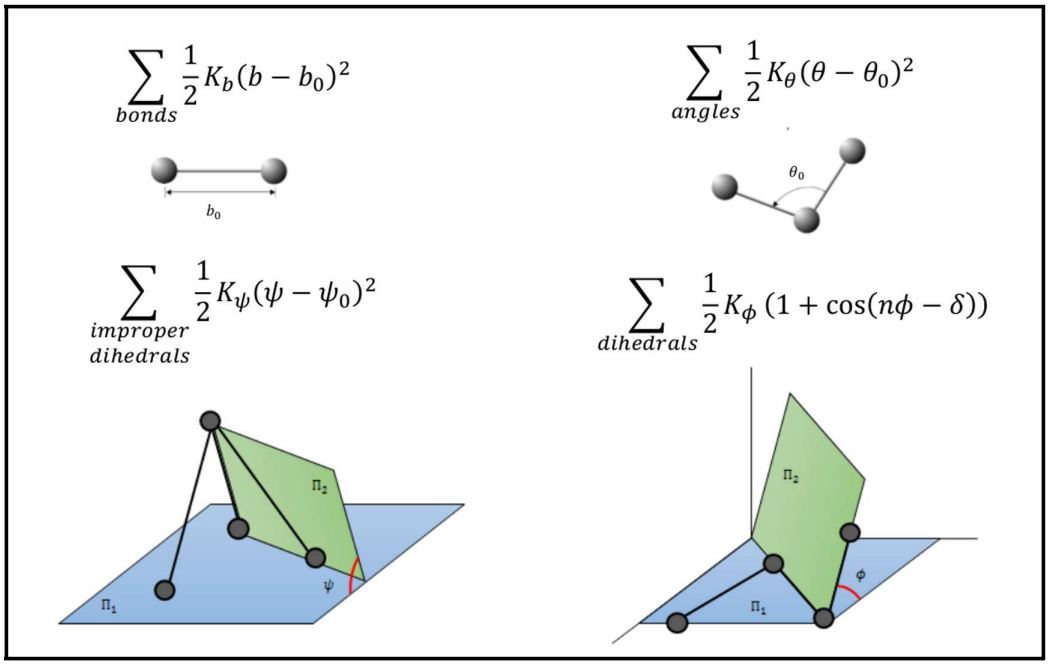


Figure 1.1: Different bonded terms appearing in force fields.

electrostatic ones:

$$V_{NB}(\mathbf{r}_1 ... \mathbf{r}_n) = \sum_{pairs(i,j)} \left[\varepsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{i,j}^{12}} - \frac{\sigma_{ij}^{6}}{r_{i,j}^{6}} \right) + \frac{q_i q_j}{4\pi\varepsilon r_{ij}} \right] \tag{1.9}$$

where ε_{ij} and σ_{ij} are the well depth and the minimum energy distance of the Lennard-Jones potential, respectively, q_i and q_j are the charges of the atoms of the pair and ε the dielectric constant. Usually, force fields store ε_i and σ_i for each atom type, and for each interacting pair of atoms i and j, ε_{ij} and σ_{ij} are obtained by applying the so-called Lorentz-Berthelot combining rules, $\sigma_{ij} = 1/2(\sigma_i + \sigma_j)$ and $\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$.

Assuming that the shape of the energy terms in Eqs. (1.8 and 1.9) and all possible parameters are stored in the force field, it is left for the software to know which atoms are interacting and the subsequent parameters needed in each time step. For that matter, MD software follows two different approaches for the bonded and non-bonded interactions. For the bonded terms, the bonded interactions that will take place throughout the simulation have to be provided to the software before the simu-

lation starts and are usually referred to as the topology of the system. Since we are not allowing bonds to break, the terms referring to angles, torsions, etc. will hold throughout the simulation, even though their value will depend on the positions of the particles and will be different at each time step, this is justified as we know that proteins or lipids form covalent bonds that are energetically very costly to break. Hence, it is mandatory to generate the topology of the system with the information of the number of bonded terms and the atom types forming these terms, determining the number of terms appearing in the sums of Eq. (1.8 and the parameters needed for each one, so that the software only has to introduce the value of the positions at each time in Eq. (1.8) to compute the bonded terms of the potential.

For the nonbonded interactions, in principle, all atoms interact with each other, making the number of interactions scale as $N(N-1)$, being N the total number of atoms (and usually a big number). To speed up the calculations, nonbonded interactions are further differentiated into short-ranged and long-ranged by a cutoff. For each atom, to establish the atoms that are within the cutoff for and to avoid computing the distance to the rest at each time step, atoms separated by a distance slightly larger than the cutoff will be stored in a list that is updated with a frequency of 10-20 steps, so only distance differences of the atoms in this list are needed for subsequent computation of the interaction potential with Eq. (1.9). For the long-range interactions, the potential is computed by means of the particle mesh Ewald summation, which scales as $Nlog(N)$ since its foundation lies in performing a Fourier transform to compute the potential through the Poisson equation in the reciprocal space of the positions.

Summarizing, we can sketch how a molecular dynamics software works: coordinates and atom type of each particle are introduced in the force field to define the potential; this potential is used to compute the coordinates and velocities of the system at the following time step in a Verlet like algorithm that moves the particles defining the new positions and therefore the new potential, repeating it for all the MD simulation steps.

This introduction to molecular dynamics is a simplified version of what actually happens inside a simulation; many subtleties followed by different softwares to reduce the computational cost and more information on the possible options allowed by MD software can be found, for example, in the GROMACS manual [23] or the NAMD's user guide [24].

1.1.3 Connection with statistical mechanics

Now that we have the time evolution of a system and the MD simulation is completed, the next step is to post process the simulation and extract information out of it, which is more difficult than expected, as in most cases it is not possible to derive conclusions just from the visualization of the obtained trajectories. On the one hand, MD can be logistically cumbersome as a trajectory generates a lot of data, namely positions and velocities for all particles, and this information has to be properly managed to derive other properties of the system. On the other hand, a theoretical background in statistical physics is required, as it is the mathematical tool needed to extract mean values of the target quantities out of the time evolution of the system. The equations of the velocity Verlet algorithm have an associated error proportional to Δt^3, which means that as the simulation goes on the error on the magnitudes computed also grow, making the outcome of the simulations not deterministic (or chaotic). We will introduce now some basic concepts of statistical mechanics that will help us understand the mathematics behind the calculation of any property and the inherent approximations lying underneath.

For each system, a phase space, Γ, can be defined, and it is formed by all plausible combinations of positions q_i and momenta p_i that a system of N particles can have:

$$\Gamma \equiv \{q_i, p_i\} \qquad i = 1...3N \tag{1.10}$$

A state is defined as a point of Γ, this states allocate one value for each coordinate and momenta, and the ensemble of all these states constitutes indeed the phase space. We assume that Γ will be static in time, this is that no external forces or time-dependent perturbations will be introduced. Each state will have an associated probability of materializing defined by the probability density function $\rho(q, p)$ in the phase space:

$$\rho(\{q_i\}, \{p_i\}\,) \equiv \rho(q, p) \tag{1.11}$$

so that:

$$\rho(\{q_i\}, \{p_i\}) dq_1 ... dq_{3N} dp_1 ... dp_{3N} \equiv \rho(q, p) dq dp \tag{1.12}$$

is the probability that the system lies in an infinitesimal volume of phase space. In biology, a space smaller than Γ is usually defined as there are some states that will not be physically accessible and we usually do not include them in our simulations. An example of those states can be the ones that have covalent bond breaks, mathemati-

cally we can say that we set the probability of these states to zero; this new smaller space is usually called configurational space.

The combination of all points in the phase space plus the probability density function contains all the information of the system and it is called ensemble. If our system is at thermal equilibrium, that is, it has a constant number of particles, volume and temperature (N,V,T), it is said to be in the canonical ensemble and its probability density function is defined as $\rho = e^{-H(\{q_i\},\{p_i\})/k_B T}$, where k_B is the Boltzmann constant and T the temperature and $H(\{q_i\},\{p_i\}) \equiv H(q,p)$ is the Hamiltonian of the system $H(q,p) = K(q,p) + V(q,p)$, which is the sum of the kinetic term $K(q,p)$ and $V(q,p)$. With these definitions we can write the first postulate of statistical physics, the mean value of any magnitude that depends on the coordinates and momenta is defined as:

$$\langle A \rangle = \frac{1}{Z} \int \rho(q,p) A(q,p)\, dq\, dp \tag{1.13}$$

Where we have normalized the value by dividing it by the partition function (Z), which is the integral of the probability density throughout all phase space. The idea behind Eq. (1.13) is that in order to know the mean value of a quantity, we need to compute it at each point of the phase space and weight it with the probability associated to the appearance of that state, as it must be considered that there are high energy states that will not contribute to the average. This integral is in general not feasible, since the number of elements in the ensemble is supposed to be big and the number of integration variables explodes with the number of particles. However, MD simulations follow another approach to evaluate the mean values, this is, if we let the system evolve a sufficiently large time it will be able to explore all the phase space yielding the time average:

$$\bar{A} = \lim_{t\to\infty} \frac{1}{t} \int_0^t A(q(t),p(t))\, dt \tag{1.14}$$

Furthermore, under these circumstances the probability of a state $P(q,p)$ will be the fraction of time that the system spends in that state, meaning that we can even substitute the partition function by this probability obtained from the simulations, yielding the free energy F:

$$\rho(q,p)dpdq = P(q,p)dpdq \Rightarrow F = -k_B T ln(P) \tag{1.15}$$

And projecting the free energies on some coordinates can lead to free energy profiles that can describe the energetic barriers separating some events of interest. Finally, a system is said to be ergodic if it satisfies that the time average is equal to the mean value:

$$\langle A \rangle = \bar{A} \tag{1.16}$$

As we are not able to compute ensemble averages using Eq. (1.13), in MD simulations we have to assume that the system is ergodic and the time average will arise reliable results. This hypothesis lies under all molecular dynamics simulations and it is very tricky to deal with because it is not possible to know beforehand if this assumption is acceptable. This is the reason why molecular dynamics software's manuals will not specify how long you have to run your simulation in order to estimate physical quantities, it will depend on the size of the portion of interest of the configurational space and the time scale needed by the system to explore it. So what can we do to validate the simulations? In general, there are a few rules that molecular dynamics works follow, like doing different replicas of the same simulation, or checking if the computed magnitude is robust over the simulation time, but what usually works best is having some experimental data that somehow validate your results.

To conclude this part of the introduction, we will define a couple of magnitudes that will be computed repeatedly in almost all simulations of this thesis, as they serve as indicators of what is actually happening in a simulation without the need of carefully visualizing all frames, which can be really cumbersome in some situations. They will also serve as examples of the previously commented statistical quantities.

1.1.4 Root Mean Square Deviation

The first indicator is called Root Mean Square Deviation (RMSD), and consists in computing the average deviation of the coordinates of a set of atoms $\mathbf{r}_i(t)$ at each time t with respect to a reference structure $\mathbf{r_{ref}}$:

$$RMSD(t) = \sqrt{\frac{1}{N}\sum_{i}^{N}(\mathbf{r}_i(t) - \mathbf{r_{ref}})^2} \tag{1.17}$$

So that the average deviation of the structure throughout the simulation time can be plotted, usually, solid rigid rotations and translations are subtracted so that if our RMSD has steep increases or major changes through time will mean that our system

is changing and something is happening, however a plateau in general will mean that the simulation is stable through time as no conformational changes are taking place.

1.1.5 Root Mean Square Fluctuation

The root mean square fluctuation (RMSF) is the time average of the RMSD, so it yields a number per set of atoms, instead of a time series. Equivalently, it is the standard deviation of the position of a set of atoms. In this thesis, this magnitude is usually presented as a plot of its value for each amino acid of the sequence, so it can be observed that amino acids belonging to α-helices have less mobility than those belonging to non structured domains. Notice that again solid state movements are also subtracted from the RMSF.

1.2 Potassium channels

1.2.1 Biological introduction

This work will focus on potassium channels belonging to the families KCNN and KCNQ, so in the following they will be briefly introduced from a biological perspective, to characterize them, or at least to have a perspective of the biology involved. We will try to answer the following questions: how are they synthesized?, where are they located? and what is their function? But before we do that we will make a brief introduction to proteins.

Proteins are polypeptide chains; this means that they are formed by several amino acids joined by peptide bonds. These are formed when the carboxyl group of one amino acid covalently bonds to the amino group of the following amino acid, losing in the process a water molecule. The chain is elongated repeating this process until the whole protein is formed. As a consequence, the first amino acid keeps its amino group intact and the last one its carboxyl group. Therefore the chain is said to extend from its amino terminus (N terminus) to its carboxyl terminus (C terminus). This polypeptide chain is usually refered to as the sum of the backbone and the side chains, being the backbone the succession of the $NH - C_\alpha H - C' = O$, where the NH is the remain of the amino group and $H - C' = O$ of the acid one after the peptide bond formation, and the C_α is the alpha carbon of the amino acid. To this carbon the amino acid side chains are attached. The 20 amino acids that appear in proteins (also known as canonical) are distinguished by its side chain.

Potassium channels

The structure of proteins can be very intricate and it is usually separated in four terms. The primary structure takes into account the sequence of the protein, the secondary structure considers the local geometric arrangement of the protein, which can be ordered (alpha helices or beta sheets) or disordered (loops, turns or bend), both alpha helices and beta sheets are stabilized by hydrogen bonds. The tertiary structure refers to the assembly of several secondary structures to form domains, which are regions that are able to perform chemical or physical tasks. Finally, quaternary structure considers the assembly of different tertiary domains [25].

Synthesis and folding

Protein synthesis is a core process in each living cell as it consists in the creation of proteins, the biological entities that perform all the tasks needed for the proper functioning of the cells. The whole synthesis process is very intricate and here we will only introduce the basics to follow the results obtained in this thesis. The information required for the synthesis lies in the genes and it is stored in a polymer called deoxyribonucleic acid (DNA). All information is stored in a four-letter code, made by four nucleotides abbreviated by A, T, C and G. To carry the information stored in the DNA to the ribosomes, the macromolecular machines that perform the synthesis, a process called transcription is followed, it consists in replicating the information stored in the double-stranded DNA to a single-stranded ribonucleic acid (RNA), which travels through the cell until it is eventually found by the ribosome. Inside the ribosome the next step of protein synthesis, called translation takes place. It is the process by which the ribosome reads the information carried by the RNA to assemble a sequence of amino acids that will become a protein [26]. While translation takes place, the resulting nascent peptide chain exits through the ribosomal exit tunnel into the cytosol within the cell. The addition of amino acids to the nascent chain is a slow process considering that it lies in the nanoscale; the synthesis rate is around 6 amino acids per second in eukaryotic cells and around 20 in *E.coli* bacteria [27].

The last step of the synthesis is the folding of the nascent chain into the wild type structure, this process starts in the ribosomal exit tunnel and takes place vectorially from the N to the C terminus, it has been shown that the appearance of small α-helices inside the last part of the tunnel, also known as vestibule, is allowed, while the formation of tertiary structures is prevented; these peculiarities of the ribosome in the first steps of the folding process of a protein might be of critical importance [28].

Ion channels are membrane proteins

All living cells are enveloped by a membrane, whose main purpose is to prevent diffusion between the exterior environment and the cytoplasm, as both environments have an aqueous nature, this is achieved by constituting a hydrophobic area, usually by the assembly of many lipids into a bilayer. Lipidic architecture can be diverse, but in the case of the membranes that we will simulate in this thesis, it will consist in a polar phosphate group from which two hydrophobic chains come out, so that the hydrophilic heads face the water and the chains of the two layers constitute the hydrophobic surface. Membrane proteins can be classified in two main groups: integral and peripheral; integral proteins have one or more segments embedded in the membrane, a subset of these are the so called transmembrane proteins, which have domains that span the membrane and have amino acids both in the extracellular and intracellular aqueous regions, as it is the case of ion channels [26]. On the other hand, peripheral proteins are bound to the membrane indirectly by integral proteins or directly by interactions with the polar heads of the phospolipids. As a proof of the importance of membrane proteins in cells, it has been stated that they comprise nearly 30% of all proteins in eukaryotic cells [9].

Therefore, following their synthesis, ion channels have to travel to the membrane where they have to be located in a very particular conformation so that it forms a pore. The channels studied in this thesis belong to the KCNQ and KCNN families, which form the pore as a result of the tetramerization of four identical subunits, so before even addressing the functioning of ion channels, the complexity of the process by which they are synthesized and inserted into the membrane is outstanding.

Function of ionic channels

The cytosol or intracellular fluid is composed by a large variety of macromolecules and smaller organic molecules, together with potassium ions and aminoacids, and also, to a lesser extent, by sodium and chloride ions solvated in water. On the other hand, the extracellular medium has a high concentration of sodium and chloride ions [26].

The difference in ion concentration has, as a result, the appearance of a difference in electrical potential between the two regions. Almost all cell membranes have an electrical potential across them, being the inside usually more negative than the outside. This potential usually has two basic functions, the first one is to provide power to the cell allowing it to behave as a battery, the second one is to allow some cells, like neurons or muscle cells, to transmit signals [26].

Potassium channels

When a cell is driven out of its relaxed state, it has two mechanisms to restore its resting potential, ion channels and ion pumps, the main difference between them is that, in general, ion pumps use energy of the cell whereas ion channels allow diffusion in favor of the ion concentration gradients. In general, ion pores are opened or closed in response to some event, this is referred to as gating. Both KCNQ and KCNN channels are sensitive to the intracellular concentration of Ca^{2+} ions, which are recognized by calmodulin (CaM), a protein known to modulate calcium in several signaling events and that is bound to the channels [26]. When opened, these channels will allow the passage of potassium ions. However, these ions are hydrated when they approach the pore, so that the channels have to dehydrate them, with this purpose, channels have amino acids next to the pore with a carboxyl group facing to the entering ions, which are capable of displace the water molecules bound to the ions. This mechanism has found to be very selective to potassium ions [29].

Regarding the purpose of KCNN and KCNQ channels in cells, they play a role in restoring the resting potential after a polarization event has been fired. Channels of both families can be found in neurons and some KCNQ channels also in cardiac tissues, therefore they have a relevant clinical significance as their malfunction can lead to encephalopathies or arrythmias when the neurons or cardiac tissues need longer times or even fail to reach their resting conditions, a more detailed description of the structures of the channels under study will be provided later on [30].

Chapter 2

Analysis of the W344R mutation

2.1 Introduction

This chapter refers to a collaboration that led to this article [1], together with subsequent work done in the same research line both to better understand the published calculations and to continue exploring the hypothesis that will be presented here. The results of this chapter comprise molecular dynamics simulations (MD), which have been introduced in the previous chapter, and some calculations using a bioinformatics software called rossetta (https://www.rosettacommons.org/). To begin with, we will give a biological introduction to the questions addressed in this work and also a small introduction of this software. Afterwards we will show and analyze the results obtained.

2.1.1 Biological introduction

The potassium channel studied in this work is the $K_V7.2$, encoded by the KCNQ2 gene. It is mainly expressed in the nervous system and it is responsible for the so-called M-current, which is critical for restoring the neuronal resting potential. It reacts to voltage changes in the cytosol by means of $Ca2^+$ concentration, which is regulated by the CALcium MODULating proteIN or calmodulin (CaM) [30], a small protein consisting of two globular domains (N- and C- lobe) that are able to carry up to two calcium ions each. Variants of this gene that suppress its function are related to several

medical conditions, such as epileptic encephalopathy [31] or benign familial neonatal epilepsy [32].

The channel $K_V7.2$ is a tetramer, which means that it is the result of the assembly of four identical subunits Fig. (2.1 A, B). Each subunit consists of a transmembrane domain, spanning helices S1 to S6 and an intracellular C-terminal domain (CTD), formed by helices hA, hB, hC and hD Fig. (2.1 D, hD is not showed). Inside the transmembrane domain there are two distinguished regions according to their functions; helices S1 to S4 comprise the so-called voltage sensor domain (VSD) Fig. (2.1 C, orange area), as there are four arginine residues [1] in helix S4 that are capable of sensing voltage differences. When triggered, the VSD will induce a structural change that will result in the opening of the pore. On the other hand, helices S5 and S6 constitute the pore domain Fig. (2.1 C, dark blue area), which contains a GYG signature (glycine-tyrosine-glycine) that makes the channel very selective to potassium ions. In the intracellular CTD, the calcium responsive domain (CRD) stands out; it is an antiparallel hairpin formed by helices hA and hB Fig. (2.1 C, grey area), which are recognized by CaM Fig. (2.1 C, red area) providing calcium sensitivity to the channel, it is also known as the calmodulin binding domain (CaMBD). Finally both helices hC and hD allow tetramerization. Helix hD and a large intrinsic disordered domain located between helices hA and hB do not appear in the Cryo-EM structure, due to their mobility, so they are not showed in Fig. (2.1).

This chapter is dedicated mainly to the study of the W344R mutation of the $K_V7.2$ channel, let us review now the features that make this particular variant interesting. To begin with, it has been found in patients with hereditary epilepsy [35, 36] providing clinical significance to this variant. Also, its location lies in the hA helix of the CRD, and more precisely, in the so-called IQ calmodulin binding motif, a sequence of amino acids that appears often in CaM targets [37]. The sequence of this motif is IQxxΘR, where x varies from one target to another and Θ is a bulky hydrophobic residue. As a matter of fact, W344 would be the one named Θ previously, making this location tightly related to CaM binding and consequently to the function of the channel. One of the most shocking results regarding the W344R variant was that it abolished the function of the channel without affecting the binding to CaM in the process [35]. This was remarkable as other mutations in the hA helix showed a nice correlation between

[1] Arginines are amino acids which have a long hydrophobic side chain with an extra positive charge in its end.

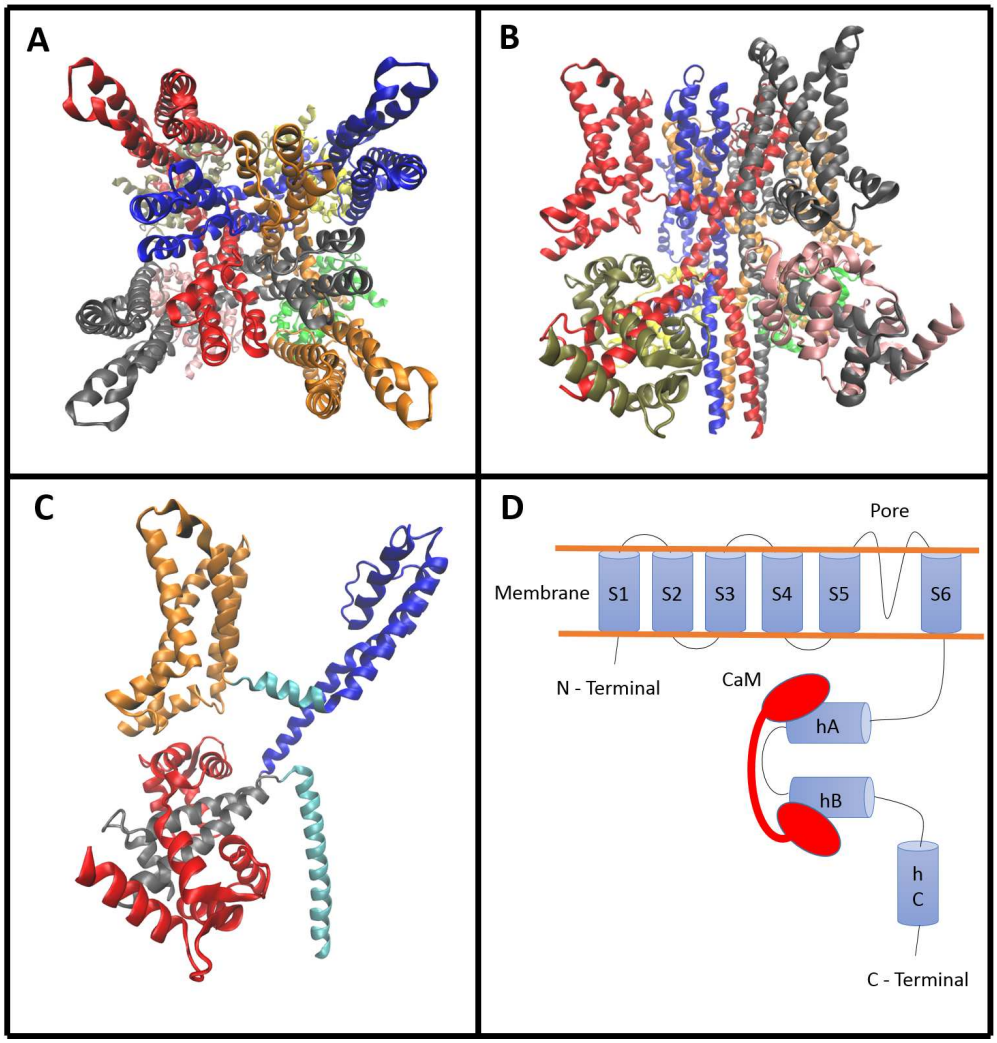


Figure 2.1: Details on the structure of the channel $K_V7.2$. **A** and **B**, top and side view of the channel respectively, where each monomer is colored differently. **C**, detail of one monomer where the voltage sensor domain or helices S1 to S4 (VSD) are colored in orange, the pore domain or helices S5 - pore helix - S6 are in dark blue. The calcium responsive domain (CRD) or helices hA - hTW - hB are in grey and calmodulin is in red. The S4-S5 linker and helix hC are left in cyan. **D**, cartoon representation of the structure of the channel, for clarity only the main helices are depicted; smaller ones like the linker S4-S5, the pore helix (between S5 and S6) and hTW (between hA and hB) are not illustrated. Figure made with VMD software [33], from the experimental structure PDB ID: 7CR3 [34].

channel function and CaM binding [38]. While the behavior of most variants suggested that the function of hA would be to serve as anchor allowing the binding to CaM, the discrepancy in the W344R mutation raised suspicions regarding the purpose of hA in the channel. Could it be that part of its functioning was missing?. In other words, what could be so deleterious to the channel function happening in hA that did not involve calmodulin? This question was the biological starting point of this work. All *in silico* calculations that will be presented in this chapter were made with the purpose of providing insight into the effects of the W344R mutation at the atomic level. Indeed, we were able to formulate a hypothesis that could explain the experimental results obtained, but before that let us review the key experimental evidences that were presented together with the computational results so that the reader can have a global perspective on this work, more details on the experiments can be found in [1].

Channel function upon W344R mutation was evaluated and the channel was found to be non functional, which was in agreement with previous results [35]. This was carried out by electrophysiology experiments, in short, bacterial cells are genetically modified to express the channel with the W344R mutation, afterwards the resulting membrane crowded with the channels is modeled as a capacitor so that membrane potential and current densities of membranes can be measured.

Furthermore, the CRD experienced a major CaM binding reduction upon the W344R mutation. This was showed by means of *in cellulo* Förster resonance energy transfer (FRET) experiments, the main idea is that cells are genetically modified to express the channel with two fluorophores linked to them, one of them will be the donor as it emits an electromagnetic radiation (blue light) and the other is the acceptor, since it also emits electromagnetic radiation (yellow light) but only if it receives the radiation transmitted by the donor, so the measurement of acceptor light is a sensor that indicates that the two fluorophores are close in space. The fluorophores used are big barrel-like structures (PDB IDs: 5OX8 - blue light [39], 3ED8 - yellow light [40]), so in order to avoid folding distortions they have to be attached at the N-terminal or C-terminal of the constructs. If we carefully design our system so that the acceptor is attached to CaM and the donor to a construct of helices hA and hB, the signals obtained will be related to CaM binding as the acceptor will only emit light if the donor is sufficiently close, considering that the CaM-hAB WT construct is stable in aqueous solution[2] (PDB ID: 6FEG [41]). The output of these experiments

[2]It was also checked that the emitted light came from CaM-CRD-like structures and it was not spurious signal from protein aggregation.

was a FRET reduction when introducing the mutation.

Therefore, it was found that the W344R mutation makes the channel nonfunctional and disrupts its binding to CaM. However, previously published *in vitro* experiments showed that this mutation does not disturb the CaM-CRD binding [35]. In fact, these experiments were repeated with the same outcome [1]. As the functional characterization is done *in cellulo*, we have experimental evidence that the mutation disrupts CaM binding (and subsequent channel function) *in vivo* but not *in vitro*.

One important difference between these kinds of experiments is how the folding occurs, in *in cellulo* or *in vivo* experiments, the whole channel or the designed construct is synthesized by cells, therefore the folding takes place vectorially and it has the physical restraint of the ribosomal tunnel. In *in vitro* experiments however, the construct is expressed in bacteria, but to recover it these cells need to be eliminated and the construct has to be solubilized in urea, a chaotropic agent. When urea is replaced by water the folding and CaM recognition is done at once without the constraints imposed by the ribosome. The results suggest that the folding pathways followed by the channel in the two experiments are different, but to test that hypothesis we have to be sure that the ribosome has an effect on the folding, which would happen if the folding starts during translation, this is, in the ribosomal vestibule.

To directly assess the effect of the mutation in the channel's cotranslational folding, FRET experiments combined with the SecM arresting peptide were carried out. SecM is a peptide sequence that is known to interact with the ribosome, detaining the protein synthesis unless some external force acts on it, pulling it out and allowing the restoration of the synthesis [42, 43], this force can be caused by cotranslational protein folding [42]. Therefore, the SecM peptide can act as a force sensor since the protein synthesis is only restored when an external force acts on it, presumably exerted by the folding of the nascent chain and steric clashes with the ribosome. A FRET experiment was designed with a construct of the CRD that is known to fold [1]. The sequence used was [FRET_donor-hA-hTW-hB-linker-SecM-FRET_acceptor], the linker length between the CRD and the SecM that will work best is not known beforehand so different lengths have to be tested. It is believed that the CRD will start to fold in the vestibule of the ribosomal tunnel, which is around 100 Å away from the peptidyl transferase center. If the linker is too short the CRD will not be able to reach the vestibule and cotranslational folding can not take place. Conversely, if the peptide is too long the CRD will be completely outside the ribosome and the force felt by the arresting peptide will be lower, then, there is a sweet spot in which

the experienced force is maximal. The linker length used varied from 18 amino acids to 50, for each linker length the fraction of folded proteins was measured by collecting the FRET signal, finding that 22 amino acids was the most efficient one. The results showed that the WT construct in presence of CaM presented a high FRET signal, conversely, mutant constructs or constructs in absence of CaM did no. Suggesting that cotranslational folding takes place, but the W344R mutation avoids it. Another interesting conclusion that can be extracted from these results is that CaM is needed for the formation of the hA-hB hairpin of the CRD, since the absence of CaM for the native sequence the FRET signal was not observed either.

Summing up, the experiments carried out indicated that the effect of the W344R mutation was to disturb the cotranslational folding preventing CaM binding and leading to a nonfunctional channel, while presenting a wild type (WT) like behavior in *in vitro* experiments. These evidences disagree with the usual conception of protein folding in which the potential energy landscape of a protein is a funnel-like structure with the folded (native) state at its bottom [44], since it looks like something happens (presumably) cotranslationally that impedes the variant to acquire its lower state conformation, leaving the mutated channel in a non productive conformation that is ultimately related to pathologies.

2.1.2 Rosetta software

In parallel to the previously introduced experiments, computational effort was made to try to rationalize the results from an atomic perspective. Some of the results were obtained with the Rosetta software, a suite for macromolecular modeling and protein structure analysis. The scope of the software is wide, evidenced by the variety of applications of the suite, ranging from the prediction of protein structures with the robetta server (`http://new.robetta.org/`) [45] to the modeling of several systems, like protein-protein [46, 47], nucleic acid-protein [48], antibodies-immune system proteins[49], membrane proteins [50] or designing new proteins [51], among other applications.

The approach followed by this software differs from the molecular dynamics one, ruled by computing the time evolution of the system; instead, it focuses the effort on evaluating the ensemble average Eq. (1.13), as stated before, this integral is not feasible, what is done in practice is to sample the conformational space by a predefined rosetta protocol to evaluate the physical plausibility of the sampled structures.

Generally, a typical rosetta protocol can be summarized as follows [52]: the input is a biomolecule in a specific conformation (or pose), that will be transformed by the application of a mover, yielding an output structure, movers can be deterministic or based on stochastic Monte Carlo methods and usually ruled by the Metropolis criterion.

The Metropolis criterion is an algorithm that computes the acceptance probability P of the new structure by energetical comparison of both the new and the original structures. If the energy of the new one is lower it automatically accepts it and assigns a probability otherwise [53]:

$$P = \begin{cases} 1, & \text{if } E_{new} < E_{orig}. \\ e^{-(E_{new}-E_{orig})/k_B T}, & \text{otherwise.} \end{cases} \tag{2.1}$$

The variety of movers or alterations that can be applied and the devoted effort made to improving the accuracy of the score functions used to estimate the energies of conformations [54] makes the software one of the most used in bioinformatics calculations.

In this chapter of the thesis we used the so-called FlexddG package [55], with the aim of computing the CaM binding affinities of some metated channel variants. In this case the protocol used by rosetta to sample the conformational space is complex (Fig 2.2), it starts by minimizing the input structure (step 1). Then, it performs the backrub protocol (step 2), which consists in introducing displacements randomly in the backbone angles in a shell of 8 Å around the mutation site. These displacements are accepted or rejected based on the Metropolis criterion with an acceptance probability of $e^{-(E_{new}-E_{orig})/1.2k_B T}$ instead of the one in Eq. (2.1), for improve the sampling. This will generate a sampling ensemble of 50,000 backrub moves. Afterwards, the protocol follows different paths for the mutant and the WT. For the WT, the side chains of residues that were previously moved are optimized with the packing protocol, this is done for each structure of the ensemble obtained with the backrub protocol (step 3, pack). For the mutant, the mutation is introduced before optimizing the structures (step 3, mutate and pack). For each structure of both WT and mutant optimized ensembles, the free energy difference of the bound complex and the separated proteins, $\Delta G = G_{complex} - G_{unbound}$ is computed by means of the Rosetta energy function 2015 (REF2015) [56]. Finally the comparison of the free energy differences of wild type and mutant yields the binding affinity $\Delta\Delta G = \Delta G_{Mutation} - \Delta G_{WT}$ (step 5), where the final value of the binding affinity will be the average of all the 50,000 structures in the ensemble. The whole protocol is repeated 50 times to improve the sampling.

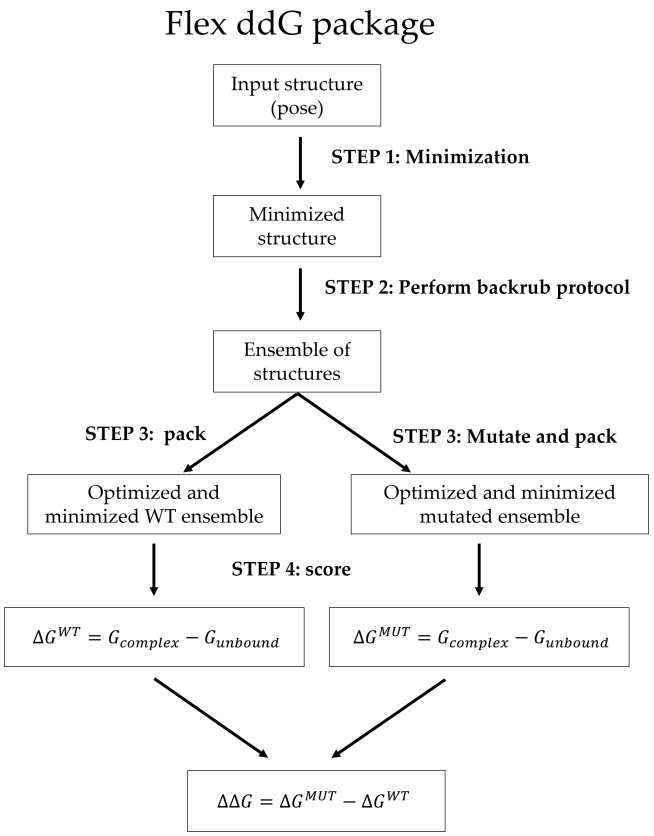


Figure 2.2: Flow chart with the protocol followed by the rosetta Flex ddG package.

In this protocol, the side chain optimization is performed with the packing protocol, the idea here is that rosetta has a list of discrete possible side chain orientations for each amino acid, known as rotamers, and the packing is the searching of the combination of rotamers of the different amino acids in the sequence that gives the lowest-energy solution, this sampling of rotamer combinations is also done stochastically. Summarizing, the FlexddG protocol provides a sampling of optimized structures that have a variety of small changes in the surrounding area of both backbone and side chains of the mutated site, so that the mutated residue can fit in a plausible conformation for the posterior score.

This protocol successfuly predicted most of the loss-of-function variants of the

KCNQ1 channel [57] and partially correlated pathogenic mutations with a destabilizing value of $\Delta\Delta G$ for the KCNQ2 channel [58]. In both cases the structure of the channels was modeled and previous versions of the score function were used.

After the rosetta FlexddG calculations have been introduced, let us make a few comments on the meaning of the computed final magnitude, $\Delta\Delta G$. A positive value of $\Delta\Delta G$ means that the mutant shows a weaker binding affinity than the wild type and the mutation is considered destabilizing. Conversely when the $\Delta\Delta G$ yields a negative value the mutation is considered stabilizing since its binding affinity is bigger than the one of the wild type. This computed binding affinity has to be considered as a qualitative estimation of the real one for a variety of reasons. The first one is that it has been found that the error associated with this technique is around 0.5 kcal/mol [59] Therefore if the absolute value of $\Delta\Delta G$ is below this quantity we will say that the mutation is neutral. It is only after the score function of 2015 that rosetta results are comparable to energy units of kcal/mol, since rosetta score functions are mixes of physical and knowledge-based terms[3] [56, 60]. The other source of error that must be considered is that we are making a quite strong assumption when conducting these calculations, it consists in assuming that the mutation will not affect the folding or the posterior assembly of the system, in this case the calmodulin binding domain of the channel and CaM, or posterior tetramerization or membrane trafficking, as we are forcing the mutation to fit in an ensemble of wild type-like conformations, by allowing only small changes in a nearby area. Even though mutations located in critical locations could affect one of the many steps needed for a protein to adopt its final wild type like conformation. Nonetheless, these calculations provide interesting insight into the mutation as they give an ensemble of possible mutant conformations that we know that could appear at least in the *in vitro* experiments, in which CaM is known to recognize the W344R variant.

[3]These terms penalize situations that are statistically infrequent, for example certain torsion and improper dihedrals [54].

2.2 Results

In this section, we will present the results appearing in reference [1] together with parallel work that keeps exploring the relevance and role of the position 344 in the channel's cotranslational folding. To begin with, in order to asses the stability of the CaM-CRD complex using a computational approach we calculated $\Delta\Delta G$ using the package Flex ddG of rosetta. The motivation for performing these calculations was twofold; first of all, it would provide us with an estimation of the binding affinity of the mutation of interest, but also with an ensemble of structures for both wild type and mutant. The computed value of $\Delta\Delta G$ for the W344R mutation was 5.1 kcal/mol, which is a typical value of a destabilizing mutation [55, 56], suggesting that the binding to CaM should not take place. To evaluate the capability of rosetta to make sensible predictions we evaluated other variants at position 344 whose impact on the channel function had been previously evaluated, namely W344E, W344F, W344G and W344Q, exhibiting a marked correlation between a high $\Delta\Delta G$ and the loss of channel function; results are depicted in Fig (2.3, A). As stated before, previous *in vitro* experiments showed that the W344R mutation is capable of binding to CaM with a slightly better affinity than the WT [35], in contrast with the destabilizing output of rosetta's $\Delta\Delta G$, this made us think that there is a stable bound conformation for the W344R variant that the conformational sampling of the rosetta protocol was failing to reach or that was being underestimated by the rosetta energy function. On its github page of materials and methods all python scripts needed for launching the calculations are provided together with the input and outputs of rosetta [61], the ensemble of structures is not provided due to storage limitations on github.

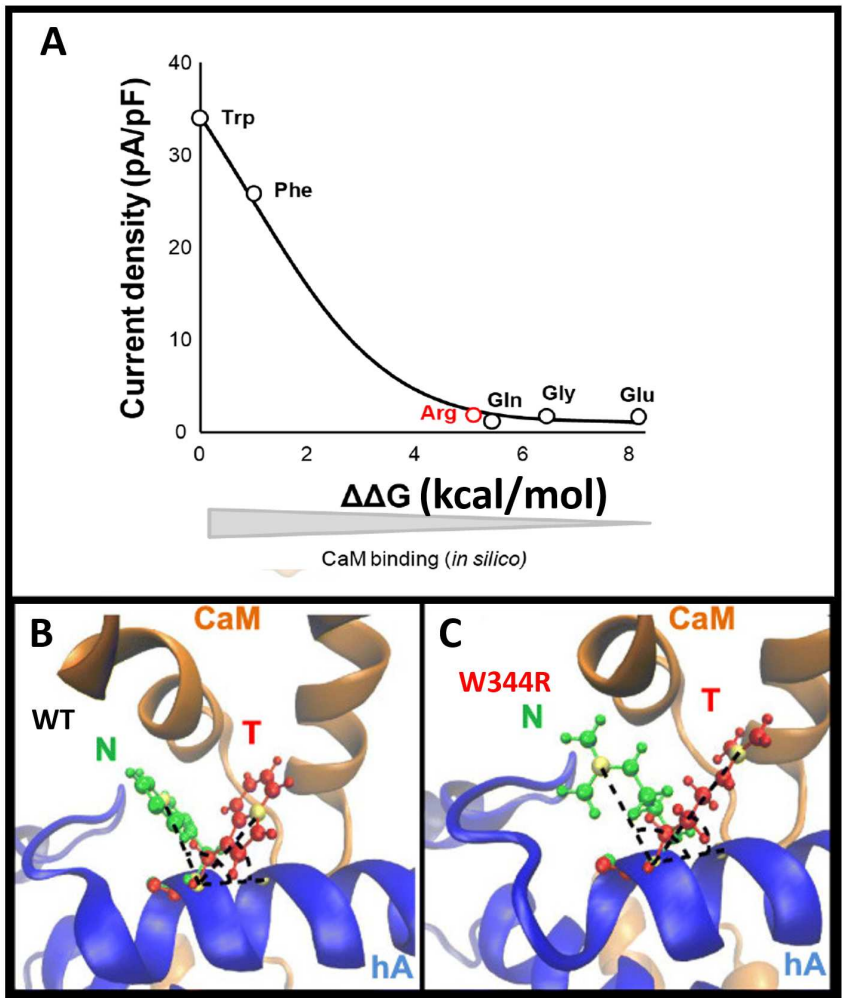


Figure 2.3: **A**, relationship between current densities and the computed binding energies of several variants of $K_V7.2$ channel, carrying the indicated mutations at position 344, from [1]. **B-C**, overlaid visualization of the native (N, green) and tilted (T, red) configurations that arose in the FlexddG calculation for WT (**B**) and W344R (**C**), with the nearby residues of CaM and helix hA. In both panels, the angle used to differentiate the conformations is plotted, the exact atoms are depicted in yellow and they are the C_α of Q341, C_α of residue 344 and $C_\delta 2$ in the case of W344 and $C\varepsilon$ in the case of W344R. Rendered with the VMD visualization software [33].

When visualizing the structures provided by rosetta, we saw that there were some populations (1.6% for WT and 3.4% for W344R) in which the side chain of residue 344 became tilted (T) towards the rest of the hA helix, instead of targeting the CaM

C lobe, as it appears in the native (N) WT structure Fig. (2.3, B and C). These two conformations are distinguished by the angles depicted in Fig. (2.3, B and C). As the binding affinity is the result of the average of the ensemble generated by rosetta, this difference in the obtained conformational ensemble could be important.

To have a better understanding of the energetic interactions of the N and the T conformations of W344R mutation with CaM, three systems were subjected to all-atom MD simulations: the WT in native conformation (WT-N), the W344R variant in native conformation (W344R-N) and the W344R variant in tilted conformation (W344R-T). All three systems were built from the rosetta output and included only the helix hA of the CRD and the C-lobe of CaM, the three are showed in Fig. (2.4, A-B). Afterwards, the systems were solvated with TIP3P water [62] in a cubic box assuring that the protein was at least 1.2 nm away from the sides, also salts concentrations of 120 mM of KCl and 5 mM of NaCl were added to mimic neuron physiology. All simulations were carried out with the NAMD 2.13 software [15] and the CHARMM36 force field [21]. The workflow followed for the three constructs consisted of 5000 steepest descent steps followed by two thermalization steps, firstly 0.5 ns in the NVT ensemble at 298 K were simulated, and secondly another 0.5 ns in NPT ensemble. In both cases temperature and pressure were kept constant using Langevin dynamics [63, 64]. Finally, 100 ns were collected again in the NPT ensemble. The length of the simulations is considered to be enough since we do not want to observe any conformational change, only a nice statistic of the interactions occurring at position 344 is sought.

Further information about these simulations can be found in its section of the github page of materials and methods [65], in particular all input files with binary files with final snapshots of the systems at each step of the workflow. Also some magnitudes computed to control the simulation are provided. Firstly, to assure that there were no major conformational changes the RMSD of all protein atoms[4] and the RMSF of each amino acid of the systems[5] are provided, together with the angles formed by the position 344[6] as defined in Fig. (2.3, B and C) so that it is clear that the studied conformations of position 344 are maintained throughout the simulation.

In Fig. (2.4, C), it is showed the mean interaction energy between some residues of the simulated system with W344, W344R-N and W344R-T. The interactions with

[4]Follow the links for RMSD of WT, W344R-T, W344R-N runs.
[5]Follow the links for RMSF of WT, W344R-T, W344R-N runs.
[6]Follow the links for the angles of WT, W344R-T, W344R-N runs.

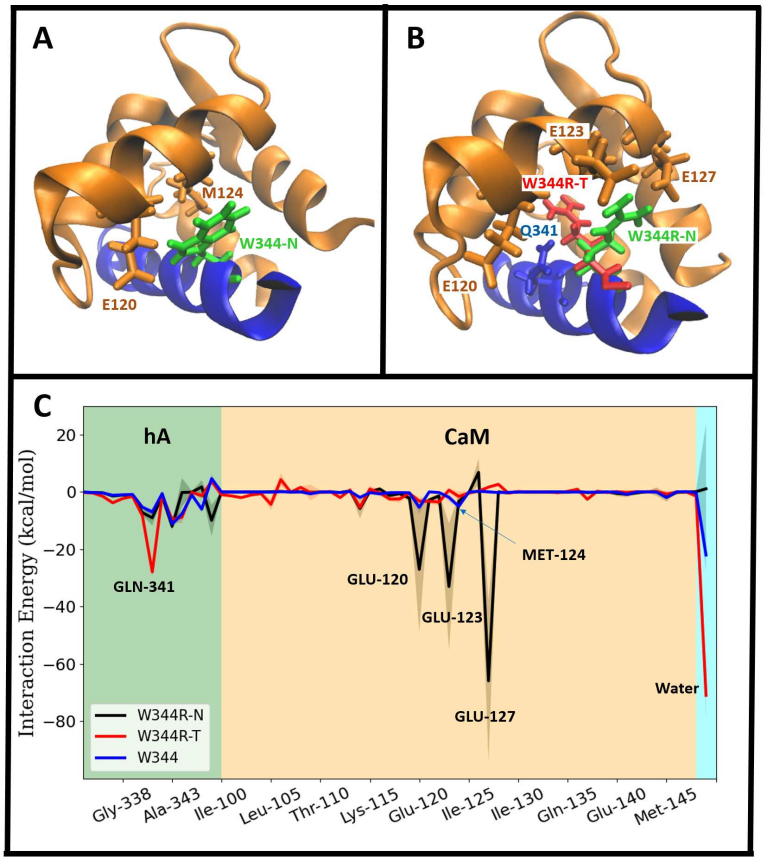


Figure 2.4: **A**, Simulated system for WT helix hA (blue), bound to CaM C-lobe (orange), where W344 exhibits a N conformation (green). Some important residues of the system are emphasized. **B**, Overlaid representation of the simulated systems containing W344R-N (green) and W344R-T (red) conformations in helix hA (blue) when bound to CaM (orange); again, some important residues of CaM and hA are depicted. **A** and **B** panels were made with VMD [33]. **C**, interaction energies of residue at position 344 for WT (blue line), W344R-N (black line) and W344R-T (red line) conformations with the rest of the amino acids in the sequence. For an easier visualization background has been colored in green for the residues of hA, orange for those of CaM C-lobe and cyan for water. Residues 329 to 332 in hA and 81 to 99 in the CaM C-lobe were also simulated but are not included in the plot as they did not interact with the position 344, lines link the mean energies obtained and errors are filled areas with the standard deviation of the interaction energies throughout the simulation.

CaM residues showed started from Ile-100 since the previous ones did not interact with position 344 in any of the three runs. Interaction energies were computed with the NAMD energy plugin [24] based on the CHARMM36 force field [21]. From Fig. (2.4,

C: blue line), we can describe the energetic fingerprint of W344; it is characterized by some attractive interactions with the nearby residues belonging to hA, and with two residues of CaM, namely Glu-120 and Met-124, that would serve as the anchor points of W344 to CaM. Finally, the mean interaction with the solvent is attractive. However, the W344R-N configuration shows much more stronger interactions with Glu-120, Glu-123 and specially with Glu-127, where the first two interactions have large standard deviations, indicating that these interactions are not maintained throughout the whole simulation, in favor of Glu-127. Notice that glutamates and arginine are negatively and positively charged residues, respectively, so their interaction compensates the charges, presumably resulting in the strong interactions observed but also in inducing a hydrophobic behavior in the arginine; once its extra charge is compensated what is left is a long hydrophobic side chain that interacts less with water, as can be seen in Fig. (2.4, C: black line), suggesting that the hydrophobic nature of the WT W344 is conserved or even enhanced by the W344R mutation when being in a N-like conformation, which would rationalize the *in vitro* binding experiments that stated that the mutation showed even better CaM binding [35]. Remember that the 344 position is located at the hydrophobic residue Θ of the IQxxΘR calmodulin binding motif so a hydrophobic nature is expected. Finally, the interactions of W344R-T with CaM are drastically reduced when compared to W344R-N Fig. (2.4, B: red line); in fact, the absence of these interactions leaves the extra positive charge of W344R-T uncompensated making it very hydrophilic, as it has a very strong interaction with water, in contrast to the smaller ones showed by W344R-N or wild type. We speculate that the hydrophilicity observed in the W344R-T conformation would prevent the binding with CaM. It also stands out the interaction of W344R-T with GLN-341 (and is also a polar amino acid), which could be the one promoting the appearance of the T configuration.

These descriptions of W344R-N and W344R-T can only be understood as a plausible explanation of how the appearance of W344R-T would be deleterious to CaM binding, as the previous comments rely on the estimation of the hydrophobicity of the amino acids by only looking at their interaction with water. This approach is naive as to properly assess hydrophobicity the entropy of the displaced water has to be considered, which is not trivial at all and is beyond the scope of this work.

To elaborate a hypothesis on the differential effect observed on the mutation upon CaM recognition, we computed which orientation is the most likely in absence of CaM, since in the end this would be the one with more probability to be encountered by

CaM during its recognition. To do so, three runs of 100 ns long of all-atom molecular dynamics of the CRD (helices hA-TW-hB, Fig. 2.5, A) of both wild type and mutant were performed. The system was built from the CRD of (PDB ID: 6FEG) structure [41]. As before, it was placed in a cubic box ensuring that the minimun distance between the protein and the box was half the cutoff (0.61 nm) so that short range interactions between periodic units are prevented. Salt concentrations of 120 mM of KCl and 5 mM of NaCl were introduced to mimic neuron physiology. The workflow followed for this set of simulations was repeated for each run, input structures were minimized by 5000 steps of the steepest descent algorithm, which was followed by two 0.5 ns long thermalization steps in the NVT and NPT ensemble, in which temperature was kept at 298 K using Langevin dynamics [64] and the pressure was kept at 1 atm also with a Langevin piston [63]. Finally, 100 ns in the NPT ensemble were collected for each of three independent replicas as production. All simulations were carried out with NAMD 2.13 [15] and with the CHARMM36 force field [21].

Further information can be found in its materials and methods github page [66], input data and snapshots of each simulation steps can be found there. The stability of the systems through the simulated time is analyzed with the RMSD[7] of all atoms of the CRD and the RMSF[8] of each amino acid. RMSD values around 0.7 nm were observed, which are rather large deviations, however when looking at the RMSF amino acids belonging to helices entailed less RMSF and visualizations of the last snapshot guarantee that the secondary structure of the systems is always maintained, even though the systems undergo large fluctuations.

Angles as defined in Fig. (2.3, B and C) were collected throughout all simulations [9] and stored to compute the free energy profile using the Weighted Histogram Analysis Method (WHAM) [67, 68] Fig. (2.5 B); an example of a time series of a simulation can be seen in Fig. (2.5 B, left inset). The projection of the free energy on the angles revealed that the WT has a potential well around the N orientation and the T configuration seems to be banned as no transitions were observed; in fact, simulations with the same workflow[10], but starting from an output structure from the rosetta protocol in which W344 was in the T configuration, showed that it quickly returned to N Fig. (2.5 B, right inset). This was repeated three times with the same outcome[11], indicat-

[7] Follow the links for RMSD of WT-1, WT-2, WT-3, W344R-1, W344R-2 and W344R-3.
[8] Follow the links for RMSF of WT-1, WT-2, WT-3, W344R-1, W344R-2 and W344R-3.
[9] Follow the links for the angles of WT-1, WT-2, WT-3, W344R-1, W344R-2 and W344R-3.
[10] Follow the links for RMSD of WT-T1, WT-T2, WT-T3 and for RMSF of WT-T1, WT-T2, WT-T3. Same comments regarding stability can be made as in previous WT and W344R simulations.
[11] Follow the links for angles of WT-T1, WT-T2, WT-T3

ing that the N configuration is the most stable one for W344. Conversely, W344R has access to a wider range of angles as the N configuration is not energetically banned, but it finds a more stable conformation around the T position. Regarding the total amount of simulated time collected in this set of simulations, it is clear that the 3 runs of 100 ns considered is long enough to observe several transitions between W344R-N and W344R-T, since the time scale associated to its energetic barrier is lower than the simulated time.

Gathering all the previous information, we can express our hypothesis that could explain the molecular mechanism according to which W344R is able to bind to CaM *in vitro* but not *in vivo*. When the folding of the CRD-CaM complex takes place at once after urea denaturation, the arginine of the W344R variant can adopt its lower state bounded conformation, that would be the N orientation, in which strong electrostatic interactions between the arginine and glutamate residues take place and the arginine exhibits a hydrophobic nature that would enhance the binding. Conversely, in the ribosomal tunnel CaM has to recognize the channel and we speculate that the cotranslational appearance of the T conformation for ariginine of the W344R mutation would prevent its proper binding to CaM, as we have showed that T orientation is the most stable one in the absence of CaM. The metastable state of the T orientation bounded to CaM has been computationally modeled, resultin in the loss of interactions with the glutamates of CaM that left the arginine as a highly hydrophilic residue, while position 344 is known to be the hydrophobic residue of the IQ-like binding motif.

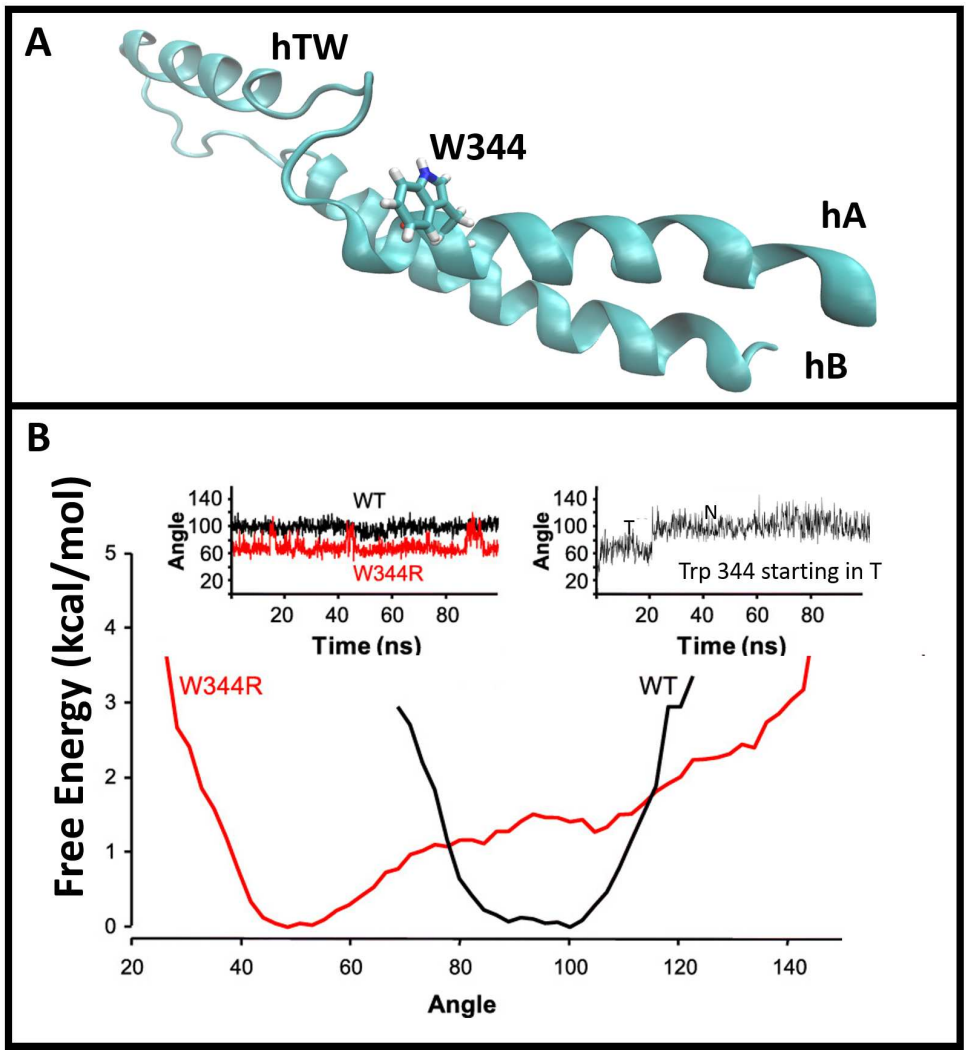


Figure 2.5: **A**, simulated system, consisting in the CRD of the $K_V7.2$ channel, and showing the location of W344, rendered with VMD [33].**B**, Data collected from simulation containing helices hA, TW, hB, free energy profile projected on the angle between the helix hA and the W344 (black) and W344R mutation (red) as defined in Fig. (2.3). The left inset refers to one of the three simulations that make up the energy profile showed. The right inset is an example of the simulations where the W344 was forced to start in a T orientation. Follow the links in the main text for the angles of the other two runs in both cases. Adapted from [1].

Part of the previous results were published in [1], now we will present further experimental and computational efforts that were made to test our hypothesis. Looking again at Fig. (2.4 C), the tilted arginine engaged in attractive interaction with Gln-341. Furthermore, the Pearson correlation coefficient of the angle of the arginine and the interaction energy between Arg-344 and Gln-341 yielded a value of 0.52:

$$corr(\theta, E) = \frac{COV(\theta, E)}{\sigma_\theta \sigma_E} = \frac{\langle(\theta - \bar{\theta})(E - \bar{E})\rangle}{\sigma_\theta \sigma_E} = 0.52 \tag{2.2}$$

Which was computed for values of time between 80 and 90 ns of the plot in Fig (2.5 B, left inset); this is near one transition between N and T. This constraining on the time used to compute the covariance was needed because otherwise the correlation was masked by the usual thermal fluctuation of the MD simulation. A value of 0.52 of the correlation means that there is a partial[12] positive correlation between angles and energies, so an increase in angle is associated with a positive increase in energy, in other words, the smaller angles of the tilted configuration are associated with more negative interaction energies, meaning that Gln-341 would help to stabilize the tilted conformation. When visualizing the T structures it looked like the arginine always lay close to that glutamine Fig. (2.4, B), which could make sense as the extra charge of the arginine could interact with the polar side chain of the glutamine. Our idea to continue working on this hypothesis of the N and T configurations was the following: mutating the glutamine could entail an increase in the N orientation probability of the mutated arginine, as we would be breaking their attractive interaction, recovering a WT-like behavior and restoring the subsequent CaM binding.

Therefore, to successfully test this hypothesis we needed a correlation between the appearance of the N configuration with experimental binding to CaM, namely, mutations in Gln-341 that result in a greater *in silico* probability associated with the N orientation would increase the experimental binding to CaM, where the latter would be experimentally verified by using an *in cellulo* FRET construct that had been able to detect CaM binding in the past [1]. It was based on placing the fluorophores at the N-terminal and C-terminal of the CRD, resulting in a sequence of [Fluorophore-hA-hTW-hB-Fluorophore], which is co-expressed with CaM, since it is known to arise FRET signal only if CaM recognizes the CRD, since it is not able to fold by itself but it is known to form a stable complex (PDB ID: 6FEG) [41].

[12]Notice that the coefficient is normalized between -1 and 1 as it is divided by the standard deviations.

Trying to anticipate energetic destabilization upon mutations on Gln-341 in CRD-CaM complex, rosetta calculations of $\Delta\Delta G$ for a few mutations were carried out. The preselection of amino acids was based on the idea that we needed amino acids whose side chain was not very big to properly fit in the hA of the CRD, aside of that a variety of them were tested. Results are in Table 2.1, rosetta predicted that all constructs would be destabilizing, but for some of the double mutants like Q341L-W344R or Q341M-W344R the global binding affinity was slightly greater than the reference single mutant W344R, indicating a partial stabilization provoked by the mutation at position 341. Further information and input and output files can be found at its supplementary information page [69].

Variant	$\Delta\Delta G$ (kcal/mol)	P_N
Q341A-W344R	8	0.24
Q341C-W344R	6	0.27
Q341I-W344R	5.1	0.48
Q341L-W344R	4.1	0.47
Q341M-W344R	4.78	0.38
Q341N-W344R	9.36	0.46
Q341P-W344R	8.65	0.26
Q341V-W344R	5.75	0.39
Q341-W344R	5.1	0.37

Table 2.1: Results of the double mutants, binding affinities computed with rosetta in kcal/mol and probabilities of the arginine to be in an N orientation.

All-atom simulations of the CRD were performed for all double mutants of Table 2.1 together with the single mutant W344R; to build the systems, the mutations were systematically introduced to the system of Fig. (2.5, A), and the same procedure as before was followed, a cubic box of TIP3P water solvent [62] was generated so that the minimun distance between the protein and the simulation box was 0.61 nm, and salt concentrations of 120 mM of KCl and 5 mM of NaCl were added to reproduce neuronal physiological conditions. Simulations were carried out with NAMD 2.13 [15] using the CHARMM36 force field [21]. For this set of simulations the previous workflow was followed, to avoid atomic clashes coming from the introduction of the mutations, 5000 steepest descent steps were carried out; afterwards, two thermalization steps of 0.5 ns in the NVT ensemble followed by 1.5 ns in the NPT ensemble were performed, temperatures were kept at 298K using Langevin dynamics [64] and pressure was kept at 1 atm also with a Langevin piston [63]. Finally, three or four replicas for each

construct in the NPT ensemble were performed to collect 450 ns as production. This time more time was simulated for each of the variants since the differences between different double mutants would presumably be more subtle than in the previous case, in which the comparison was between the W344R single mutant and the WT; more simulated time is translated into a greater resolution in the energy barrier separating the W344R-N and W344R-T states as more transitions are obtained. Similarly to the previous sets of simulations, there is supplementary information available on its materials and methods github page [70] with the usual supplementary information for each replica of each double mutant; inputs and snapshots of each step of the workflow and. For the production runs, RMSD of the protein, RMSF of each amino acid and the angles of position 344 as defined in Fig. (2.3, C)[13]).

Instead of computing the free energy profile like in Fig. (2.5, B), histograms of angle distribution probabilities of the arginine in W344R where obtained using WHAM equations [67, 68], in order to be fit to a two Gaussian function Fig. (2.6, black lines), each one representing one of the two possible orientations, integrating the Gaussians separately will yield the probability of the arginine of being at W344R-N and W344R-T. This methodology in the definition of the conformations was introduced to avoid arbitrariness in the computed probability associated to each state. The results are depicted in Fig. (2.6) and the probabilities of being at the N configuration P_N are collected in Table 2.1; it can be seen that there were some mutations that left the probability of being N (P_N) similar, like Q341A, Q341C or Q341I whereas others indeed made this probability higher, like Q341L, Q341M, Q341N, Q341P or Q341V, which seemed encouraging to test them experimentally. However, experiments showed that all double mutants had a similar behavior as the single W344R mutant, as no FRET signal was obtained, which means that CaM binding was not recovered, proving that this approach for testing the two-orientation mechanism for the CaM recognition of the K_V7.2 channel was not productive.

One issue with this kind of FRET experiments is that the FRET signal might not have been generated for a variety of reasons, making it difficult to extract information out of it, Gln-341 is indeed the Q of the IQ motif, so its alteration could be deleterious for the CaM binding, for example du to structural changes on the CRD that prevent CaM recognition or maybe by altering the interactions that are required for CaM binding, but even if position 341 is known to be important, mutating it was worth

[13]Instead of linking each picture, the reader is referred to the folders: Q341A-W344R(1,2,3), Q341C-W344R(1,2,3), Q341I-W344R(1,2,3), Q341L-W344R(1,2,3), Q341M-W344R(1,2,3), Q341N-W344R(1,2,3), Q341P-W344R(1,2,3), Q341V-W344R(1,2,3), Q341-W344R(1,2

a shot as CaM is known to have many targets with many different binding motifs [37]. Nevertheless, the double mutant research line was a milestone in our lab as it was the first time that computational results were used to formulate a hypothesis and design experiments to test it, as we were used to work the other way around, with experiments leading the way.

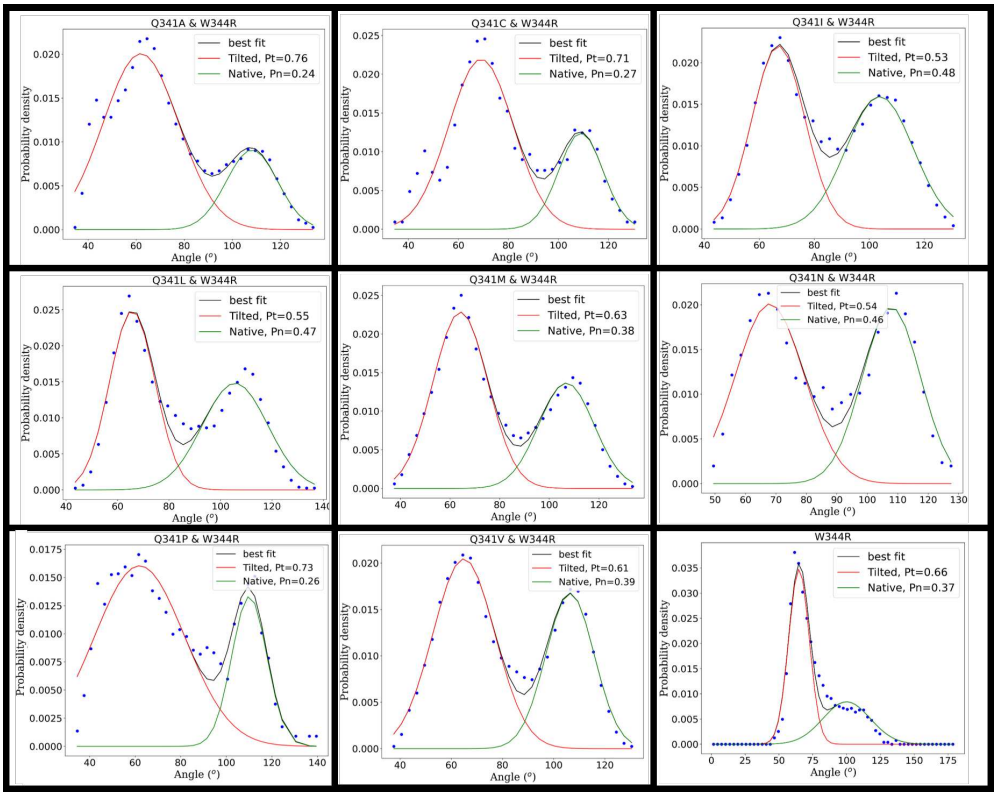


Figure 2.6: Probability densities of the angle formed by the arginine in position 344 for different mutations at position 341. Each plot is titled with the mutations included at the set of simulations, the bottom right plot is the W344R single mutant that serves as reference. Probability distributions (blue circles) have been fitted to two Gaussian distribution functions (black lines), where each Gaussian separately refers to one of the two possible states, W344R-N (green) and W344R-T (red), integration of these Gaussians yields the probability of being at the W344R-T orientation, P_T, and at the W344R-N orientation, P_N, these values are written in the legend and also in Table 2.1.

2.3 Conclusions

In this chapter, new mechanistic insight has been provided regarding the W344R mutation on the $K_V7.2$ potassium channel. The problem arose from the discrepancy of the outcomes of CaM binding of this variant in *in vitro* and in *in vivo* experiments, the experiments carried out suggested that something happened cotranslationally that prevented CaM binding upon the W344R mutation. However, since the experiments are indirect, they failed in providing mechanistic insight on what was happening. In order to assess this problem from an atomistic perspective, MD simulations together with rosetta calculations were performed. In fact, both methods follow different approaches to sample the configurational space, MD simulations rely on the time evolution while rosetta determines the plausibility of an ensemble of conformations; however, it has to be underlined that neither of them explicitly considered the cotranslational folding of the channel, since it is extremely challenging. It is noteworthy that in general, binding affinities computed through the rosetta FlexddG package and FRET experiments agreed, as it is showed in Fig. (2.3) and in the double-mutant experiments, for which rosetta predicted a destabilizing behavior and FRET experiments showed that there were no CaM binding. The exception was the *in vitro* assay of the W344R mutant, for which a destabilizing binding affinity did not correlate with the CaM binding. We think that this could be because rosetta was energetically underestimating the W344R-N conformation.

The hypothesis formulated from the obtained data is that the cotranslational effect of the W344R mutation would be to favor the W344R-T conformation, which would prevent CaM binding. Conversely, on *in vitro* experiments the W344R-N state would be the dominant one, even enhancing CaM binding. This hypothesis lies on the fact that when the CRD is immersed in water, the W344R variant has a higher probability of being at W344R-T state than W344R-N one, while the wild type always stays in the N configuration. Also, it has been showed herein that the strong interactions of W344R-T with water would infer a high hydrophilicity in a pocket that is known to be hydrophobic, whereas charge compensation upon binding of W344R-N with some glutamates of CaM might infer a hydrophobic behavior on W344R-N.

Further work on this hypothesis turned out to be unproductive; from the simulations it was clear that W344R-T was stabilized by the amino acid Gln-341, but its mutation to other amino acids did not result in an improvement in the experimental CaM binding, even though some of the mutations on Q341 promoted the appearance

of the simulated W344R-N configuration. Therefore, it was not possible to establish any relation between the appearance of the simulated W344R-N configuration and experimental CaM binding; however, it could be that the alteration of amino acid Gln-341 could be very harmful for the binding, since it is the Q of the so-called IQ motif. Nevertheless, this chapter contains one of the most important projects for this thesis since these were the first results obtained; we were able to elaborate a hypothesis that could explain the experimental results and after that even design a set of experiments to further test this hypothesis.

To convert this hypothesis in actual results there is still a long way to go, since there are many holes in the emerging picture coming from our results. First of all, it is not clear how CaM recognizes the channel, it is believed to happen cotranslationally since the hairpin formed by helices hA and hB is not stable in the absence of CaM [1], but it is not known if helix hA is already formed when CaM recognizes the channel or if is the recognition of the channel by CaM what induces the formation of the helix, so the definitions of W344R-N and W344R-T in the absence of CaM should be taken with a pinch of salt. Nonetheless, even if the two conformations are not supported by experimental information, or if it turns out that it is not the angle formed by amino acid 344 the parameter that prevents or enhances CaM binding and it is something else, the different interactions stablished by the mutated arginine with CaM for the two conformations can be enlightening, as they explain two behaviors of this arginine that would give a possible explanation for the contradictory outcomes of *in vivo* and *in vitro* experiments. The following chapter will leave aside the W344R mutation and the channel $K_V7.2$, in favor of the small conductance calcium-activated potassium channel SK2, with the aim of further elucidating how this recognition process of CaM takes place, since SK2 channel also has a hA-like helix that serves as an anchoring point for CaM.

Chapter 3

α-helix induction in SK2 by Calmodulin

3.1 Introduction

In the previous chapter, when addressing the effect of the W344R mutation on the $K_V7.2$ channel, the experimental data suggested that this mutation affects the folding pathway of the channel. One of the most revealing experiments in that matter was the force profile analysis [1], in which an arresting peptide needs to experience a force in order to resume the synthesis of the channel. In these experiments the magnitude measured was a FRET signal that could only be observed if the channel was properly synthesized and folded. The data showed that upon the W344R mutation there was no FRET signal; however, in the case of the wild type, only FRET signal was observed if calmodulin (CaM) was present, indicating that CaM plays an active role in the folding of the channel at the very early stages of its folding. As discussed in [1], it is not clear if CaM recognizes an already folded emerging segment of the channel, namely the part of the hA helix that has the IQ binding motif, or if CaM induces the adoption of the α-helix exhibited by the wild type structure. This would be crucial for understanding how the folding takes place and the role of CaM in it. The following chapter is a computational effort on the study of that unknown step of the synthesis, this is, how calmodulin recognizes its targets. In fact, the main results gathered here were published [37, 71].

3.1.1 Biological introduction

Calmodulin (CaM) is a protein formed by 148 residues separated into two globular domains, named as N-lobe and C-lobe, each of them can carry up to two calcium ions in two helix-loop-helix motifs known as EF hands. CaM is the main calcium sensor in eukaryotic cells; one example of its importance in our organism is that it is one of the most conserved proteins in evolution, being identical in all vertebrates. Mutations in CaM are usually pathogenic and the term "calmodulinopathies" has been coined to denote the associated diseases [72].

Another interesting particularity of CaM is its ability to bind to a large variety of targets, most of them display two α-helical anchor points used by CaM for binding; even though their amino acid sequences differ substantially, many of them display some motifs that are repeated often [73]. As previously mentioned, there are many uncertainties in CaM's recognition process due to its unspecificity towards its targets; a mixture of induced fit and conformational selection has been proposed both theoretically and experimentally [74, 75]. The conformational selection mechanism is the one by which CaM only recognizes α-helices, whereas in the induced fit mechanism, CaM recognizes its binding motifs and induces secondary structure on them.

Among the variety of CaM targets, two of them are the previously introduced KCNN and KCNQ ion channel families, in which CaM is crucial for its proper functioning. in this chapter we will focus on the SK2 channel. Like the $K_V7.2$ channel Fig. (2.1), SK2 is a tetramer Fig. (3.1, A-B), where each subunit is composed by six transmembrane helices (S1-S6) and a C terminal cytosolic domain, composed by helices hA, hB, hC and hD Fig. (3.1, D; hC and hD are not showed). Transmembrane helices S1 to S4 form the voltage insensitive domain (ViSD) Fig. (3.1: C, orange), which is analogous to the $K_V7.2$ VSD but it has been showed that it does not respond to voltage differences [30]. Helices S5 and S6 form the pore domain Fig. (3.1: C, blue), which is connected to the ViSD by two small helical linkers (S4-S5a and S4-S5b) Fig. (3.1: C, cyan), where S4-S5a is crucial as it lies in the cytosolic domain and serves as an anchor to CaM Fig. (3.1: C, red). In the cytosolic domain, helices hA and hB form an hairpine recognized by CaM, as hA has an IQ-like binding motif, and helices hC and hD would be responsible for tetramerization, but they do not appear in the full channel cryo-EM structure [76][1]. It is very impressive that the gating mechanism can be elucidated from the cryo-EM structures: in the case that each CaM has 3 Ca^{2+}

[1]The avalaible full channel structure is the one of the family member SK4, which is very similar to SK2, in fact all the domains explained are conserved.

ions, two in the N-lobe and the other in the C-lobe, the two lobes are respectively bound to the linker S4-S5a and to hA, (PDB ID: 6CNN [76]). Conversely, if CaM loses the Ca^{2+} in the C-lobe, the N-lobe does not appear in the cryo-EM structure, suggesting that it is not longer bound to the S4-S5a linker and does not have a stable conformation (PDB ID: 6CNM [76]). Therefore, gathering the previous information, the hA helix would be the anchor point of CaM and the S4-S5a linker would serve as a lever that is activated by CaM for gating.

The motivation of this chapter came from a work published a couple of decades ago, in which NMR experiments showed an ensemble of structures of a small region of the SK2 channel (residues 396–487) spanning the hA and its CaM binding domain Fig. (3.2, A and B), showing a stable helical turn [2]. As the peptide was immersed in water without any other cofactors, these results suggest that in this case a pre-helix is required for CaM binding and the rest of the helix would be induced afterwards, since this fragment is folded in the final structure Fig. (3.2, B). A detailed analysis of the binding domain of the SK2 channel reveals that it resembles an IQ motif [37], although its sequence, LRxxWL, is different to the standard IQ, IQxxΘR, where Θ is a hydrophobic residue.

In this chapter we will study how part of the calmodulin binding domain (CaMBD) of the SK2 channel acquires its secondary structure. We will focus on the linker 421-NAAANVLRETWLIYKNT-437 Fig. (3.2 C top), which from now on will be referred to as SK2 CaMBD, and we will compare its simulated secondary structure in absence of CaM and bound to the CaM C-lobe, the role of the CaM C-lobe in those simulations will be static, meaning that the recognition process will not be simulated, as it is out of the scope of this work, so we will focus only on the behavior of the SK2 CaMBD. We have made some strong approximations, such as reducing the system to half of CaM and a small linker of the SK2 channel, and furthermore, simulationsin presence of CaM will start from a bound conformation. Even under these approximations, simple molecular dynamics are not able to capture the process of secondary structure formation as we would need to simulate unaffordable time scales. Therefore we will make use of the so-called enhanced sampling techniques. In the following we will give a short introduction of the ones used in this chapter.

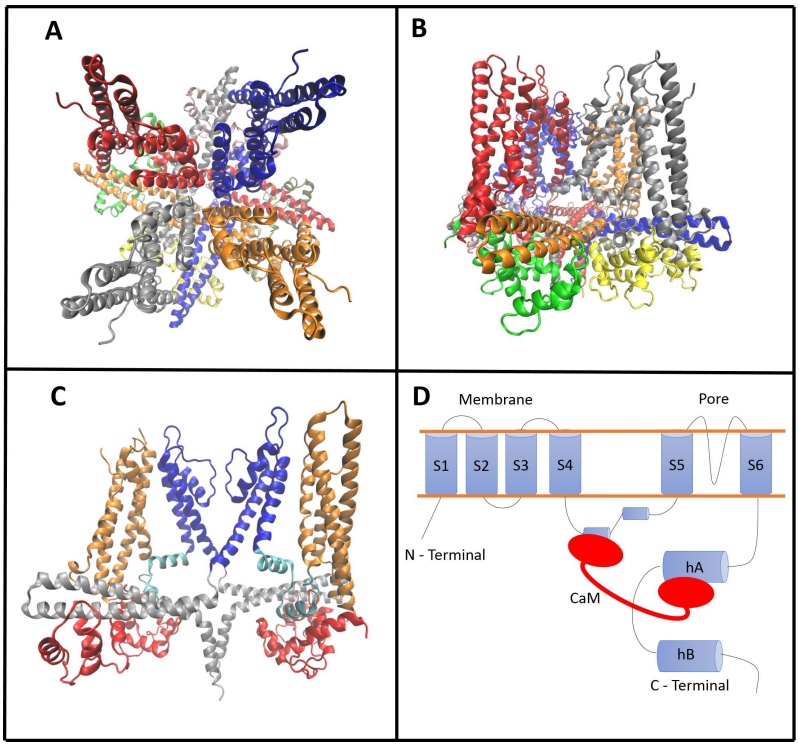


Figure 3.1: Homology model of the SK2 channel, generated with Swiss model (https://swissmodel.expasy.org/) and represented with VMD [33, 77, 78]. **A** and **B**, top and side view of the full channel, respectively, where each monomer and calmodulin are coloured differently. **C**, details of two opposing monomers, in which helices S1-S4 or Voltage insensitive Domain (ViSD) are in orange, linkers S4-S5a and S4-S5b are in cyan, pore domain or helices S5-S6 are in blue, helices hA and hB or Calcium Responsive Domain (CRD) are in grey and calmodulins in red. Notice that the binding of CaM to helix hA can be observed in the left monomer, whereas its binding to S4-S5a is apparent in the right monomer. **D**, cartoon representation of the structure of the channel.

3.1.2 Enhanced sampling techniques

As stated in the introduction of this thesis, molecular dynamics (MD) simulations usually consist in performing ensemble averages in order to compute mean values of quantities, or capturing a dynamical transition between two states by computing the free energy barrier of the event. In both cases it is crucial to achieve a good sampling of the conformational space to obtain reliable results. In other words, we need our system to be ergodic in an affordable time scale. In general, this is very challenging for biological systems, even for the most simple ones, as they have very complicated potential landscapes with several local minimums separated by high energy barriers. To help the system to overcome these energetic barriers more easily and achieve a more efficient sampling, several methods have been proposed under the name of enhanced sampling techniques [79]. In this chapter we will make use of two of them, namely Replica Exchange with Solute Scaling (REST2) [80] and metadynamics [81]. We will make a give introduction to the basic features of both methods before moving on to the results section.

REST2

The REST2 technique lies in the more general framework of the replica exchange methods. This general framework consists in altering the simulated ensemble; instead of having one simulation at one temperature in the canonical (NVT) or isothermal-isobaric (NPT) ensemble as usual, the simulation will be performed in a generalized ensemble (X), which will consist of M non-interacting replicas $(x_1, \dots, x_M)$ where each x_i is a set of coordinates and momenta of the N interacting particles of one canonical or isothermal-isobaric ensemble $(q_i^1, \dots q_i^N, p_i^1, \dots p_i^N)$, weighted by its own Boltzmann weight factor:

$$e^{-\beta_i H_i(q_i^1 \dots q_i^N, p_i^1 \dots p_i^N)} \tag{3.1}$$

where $\beta_i = 1/k_B T_i$. Since the replicas are not interacting, the Hamiltonian of the generalized ensemble is separable, making the weight factor of the generalized ensemble ($W_{GE}(X)$) also separable:

$$W_{GE}(X) = e^{-\beta H(q_1^1, \dots, q_M^N, p_1^1, \dots, p_M^N)} = \exp\left\{ -\sum_{i=1}^{M} \beta_i H_i(q_i^1, \dots, q_i^N, p_i^1, \dots, p_i^N) \right\} \tag{3.2}$$

And each Hamiltonian is defined as the sum of the kinetic and potential energies, $H_i(q_i^1 ... q_i^N, p_i^1 ... p_i^N) = K(p_i^1 ... p_i^N) + V(q_i^1 ... q_i^N)$. We want each replica to perform a random walk in this generalized ensemble, and this is achieved by exchanging replicas occasionally. When these exchanges occur, for example between replicas i and j, the exchange can be written as follows:

$$X = (x_1, ..., x_i, ..., x_j, ..., x_M) \rightarrow X' = (x_1, ..., x_j, ..., x_i, ..., x_M) \tag{3.3}$$

Which means that the coordinates and momenta of the particles of replicas i and j are swapped. These exchanges will take place with a probability $\omega(X \rightarrow X')$, we will impose that the average number of accepted trial moves that result in the system moving from state X to X' is the same that results in moving from state X' to X, so that the system is not biased towards a specific state, this is called the detailed balance condition and has the following mathematical form:

$$W_{GE}(X) w(X \rightarrow X') = W_{GE}(X') w(X' \rightarrow X) \tag{3.4}$$

A system that satisfies this condition is said to be a reversible Markov chain and performs a random walk in the generalized ensemble. Rearranging the terms we can compute the ratio between $w(X \rightarrow X')$ and $w(X' \rightarrow X)$, using (3.2):

$$\frac{w(X \rightarrow X')}{w(X' \rightarrow X)} = \frac{W_{GE}(X')}{W_{GE}(X)} = \frac{\cancel{e^{-\beta_1 H_1}} ... \, e^{-\beta_i H_j} ... \, e^{-\beta_j H_i} \, ... \cancel{e^{-\beta_M H_M}}}{\cancel{e^{-\beta_1 H_1}} ... \, e^{-\beta_i H_i} ... \, e^{-\beta_j H_j} \, ... \cancel{e^{-\beta_M H_M}}} \tag{3.5}$$

All the exponential terms that do not belong to i and j and the kinetic terms also cancel out, yielding:

$$\frac{w(X \rightarrow X')}{w(X' \rightarrow X)} = e^{[\beta_j - \beta_i](V_i(q_i^1 .. q_i^N) - V_j(q_j^1 ... q_j^N))} = e^{-\Delta} \tag{3.6}$$

To define the exchange probability between two replicas we have the ratio between the two transition probabilities plus the obvious condition that probabilities must be positive and less than or equal to one. This two constraints are not enough to define $w(X' \rightarrow X)$ and many choices can be made, one of them is the Metropolis criterion [53]:

$$w(X \rightarrow X') = \begin{cases} 1, & \Delta \leq 0 \\ e^{-\Delta}, & \Delta > 0 \end{cases} \tag{3.7}$$

Different enhanced sampling methods lie under this general framework, each of

them with its particularities. For example in the Temperature Replica Exchange Method [82] the replicas are simulated at different temperatures β_i, but with the same Hamiltonian. In this thesis the method followed is Replica Exchange with Solute Scaling (REST2), in which the replicas are simulated with different Hamiltonians, but at the same temperature. Methods that exchange Hamiltonians are usually called Hamiltonian Replica Exchange (HREX) and have the advantage that Hamiltonian functions are very flexible when selecting which terms are interchangeable. Specifically, REST2 simulations maintain all replicas at the same β_0, but the potential terms of each Hamiltonian are scaled differently, by a factor less than one, to reproduce a variety of effective temperatures β_i. The advantage over temperature replica exchange is that the potential is further separated into contributions coming from the solute (p) and the solvent (w) so that the solvent is left unscaled:

$$V_i^{REST2} = \frac{\beta_i}{\beta_0} V_i^{pp} + \sqrt{\frac{\beta_i}{\beta_0}} V_i^{pw} + V_i^{ww} \tag{3.8}$$

Where the square root in the interface between solute and solvent terms arises naturally from the Lennard-Jones combination rules. Applying the detail balanced condition to this scaled potential yields:

$$\Delta_{ij}^{REST2} = (\beta_i - \beta_j) \left[(V_j^{pp} - V_i^{pp}) + \frac{\sqrt{\beta_0}}{\sqrt{\beta_i} + \sqrt{\beta_j}} (V_j^{pw} - V_i^{pw}) \right] \tag{3.9}$$

Where we recall that the exponential of this quantity rules the exchanges by the Metropolis criterion, which is therefore very sensitive to potential energy differences of solutes and their interface with the solvent for the different replicas. Notice that the solvent-solvent interactions, which are the most abundant ones and, therefore, introduce the biggest energy fluctuations, does not contribute. Nonetheless, only exchanges with neighboring replicas are allowed as in practice non-adjacent replicas will have very low exchange probability.

The REST2 method has been implemented to go one step beyond as it allows the user to further divide the solute into a "hot" and a "cold" region, so that only the hot region and its interface with the cold one or the solvent will contribute to the exchanges between replicas. This is done to substantially decrease the number of replicas needed to have an acceptable exchange probability in the effective temperature range of interest, since the number of replicas needed to obtain efficient sampling scales with the square root of the total degrees of freedom of the terms contributing to the

exchanges [80].

Since only the first replica, the one that has a Hamiltonian that is not scaled, will yield physical results, the dynamical information of the system will be collected at this first replica. However this will have conformations that have traveled through all the variety of Hamiltonians, where the ones weighted by smaller values will have less energy barriers between conformations so it will be much more probable that these structures present higher energy conformations, considerably enhancing the conformational sampling, as the appearance of these structures in the usual NVT or NPT ensemble is much lower since the systems will tend to get stuck in local minima.

Metadynamics

The approach followed in metadynamics simulations is different [83], in this case we will accelerate the sampling of a collective variable S, that is a function of the coordinates of the system $S = S(\mathbf{R})$, in such a manner that this variable fully captures the dynamic process under study. For simplicity, the collective variable will be one-dimensional and therefore the free energy profile will be projected onto a 2D plot. Once the collective variable is set, a history-depending bias potential $V_G(S,t)$ will be added, normally this potential is constructed by adding Gaussian functions centered around the values of the collective variable that the system has at a given time t', $S(\mathbf{R}(t'))$:

$$V_G(S,t) = \int_0^t \omega\, e^{-\left(\frac{S(\mathbf{R})-S(\mathbf{R}(t'))}{2\sigma}\right)^2} dt' \tag{3.10}$$

where the constant ω is the height of the Gaussian function and σ its width. The bias potential $V_G(S,t)$ is added to the Hamiltonian of our system, preventing the system to stay at the value of the collective variable $S(\mathbf{R})$. These Gaussians will be continually filling the minima of the energetic landscape projected on $S(\mathbf{R})$, making the free energy barriers smaller until at some point all the local minima are full, yielding that:

$$V_G(S, t \to \infty) = -F(S) + C \tag{3.11}$$

Where C is an additive constant and $F(S)$ is the projection of the free energy of the unbiased system on the collective variable. Under these circumstances, the effective potential felt by the particle will be a constant and therefore the probability distribution of the states alongside the collective variable will also be a constant, and the characterization of the unbiased free energy profile using (Eq. 3.11) will be con-

verged. However, when running a simulation it is difficult to know when convergence is achieved to stop our simulation. In practice, it might happen that we continue adding Gaussian functions that introduce errors, making the correct value of $F(S)$ fluctuate with fluctuation values depending on the size of the added Gaussians. To alleviate this, the Well-Tempered metadynamics method was introduced [84], in which the height of the added Gaussians is scaled by w:

$$w = \omega e^{-V_G(S,t)/\Delta T} \tau_G \tag{3.12}$$

Where ω is the bias deposition rate, $V_G(S,t)$ is the value of the bias potential at the value of the collective variable S at time t, ΔT is a temperature-like value that can be tuned together with ω to control how fast the height of the added Gaussians decreases and τ_G is the Gaussians deposition rate. With this scaling scheme the final value of the free energy has to be scaled by a constant depending on the temperature of the simulation and on ΔT:

$$V_G(S, t \to \infty) = -\frac{\Delta T}{T + \Delta T} F(S) + C \tag{3.13}$$

Note that at the beginning of the simulation, since the height of added Gaussians will have bigger values, it allows the bias potential to quickly fill the minima of the potential and enhancing the conformational space sampling, but as the filling goes on the scaling will make the introduced potential decrease and eventually converge to an accurate value of the free energy alongside the collective variable.

The choice of the collective variable is crucial when setting up a metadynamics simulation. Indeed, guidance and recommendations on how to choose them has been extensively reviewed [83, 85]. Essentially, metadynamics simulations are only recommended when a set of collective variables can describe a process unambiguously. Let us consider an illustrative example: if we want to measure the free energy of the dissociation of two molecules, which would play the role of a lock and a key, we could think that the distance between the center of mass could be a good collective variable and that we could project the free energy on the axes of the lock to nicely describe the dissociation process. However, for a given distance between the two center of masses the relative orientation of the molecules can differ, so that for small distances we would have a very stable configuration in which lock and key would be aligned and several others in which a rotated key is clashing with the lock, making the convergence of this simulation very cumbersome.

3.2 Results

This chapter has two goals, the first one is to reproduce the experimental results in which a helical turn of the SK2 CaMBD was observed by NMR experiments in aqueous solution [2], as it would be the type of conformation that CaM would encounter when recognizing the channel, and the second one is to study how this part of the channel obtains its final α-helical conformation after it has been recognized by CaM.

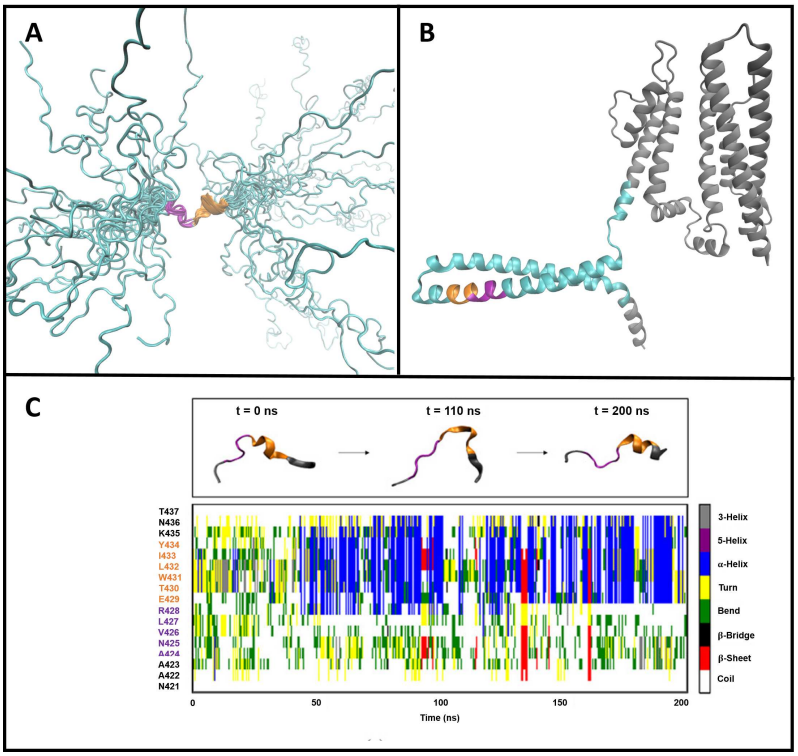


Figure 3.2: **A**, NMR ensemble of structures from PDB ID: 1KKD [2]. Residues forming an α-helix are colored in orange and ordered but not helical structures in purple. **B**, localization of residues in **A** in a monomer of the channel, with the same color code. **C**, three snapshots of the REST2 simulations of the N421 to T437 fragment immersed in water (top) and secondary structure of the lowest temperature replica throughout the simulation (bottom), according to the DSSP secondary structure definition [86]. Adapted from [37]

To asses the first goal, we tested the stability of the pre-helix in water Fig. (3.2, A and B), for that purpose we ran REST2 simulations of the 421-NAAANVLRETWLIYKNT-437 fragment from (PDB ID: 1KKD [2]), which belongs to the CaMBD and contains

the pre-helix immersed in water [37]. All simulations were performed in GROMACS 2016.4 [16] patched with PLUMED 2.4.2 [87] and using the CHARMM36 force field [21]. To build the system, we solvated a dodecahedral box with TIP3P water molecules [62], with a minimun distance between protein and walls of 1.5 nm. K^+ and Cl^- ions were added to achieve a physiological KCl concentration of 0.15 M. For the REST2 simulations, 15 replicas were set spanning a range of effective temperatures from 298 K to 373 K, where the temperature of replica i (i from 0 to 14) is given by $T_i = T_0(\frac{T_{14}}{T_0})^{i/14}$, so that the temperature difference between replicas smoothly increases with i. The hot subsystem, in which the Hamiltonian is scaled, is constituted by the peptide and ions to neutralize its charge, and the cold one, in which the same Hamiltonian is used for all replicas, by the solvent. The workflow followed in these REST2 simulations for each replica is the following. To begin with, a steepest descent energy minimization was done until the maximum force was less than 100.0 kJ/(mol·nm), afterwards, 100 ps of NVT equilibration followed by another 100 ps of NPT equilibration were carried out with restrictions on the protein's heavy atoms, temperatures were kept at 298 K using a velocity-rescaling thermostat [88] and pressures at 1 bar using the Parrinello-Rahman barostat [89] in the NPT thermalization. Finally, 200 ns in the NPT ensemble were collected as production for each replica, with exchanges attempted every 100 ps. The input parameters and a snapshot of the first replica at each time step can be found at its materials and methods github page. In the case of the production run, time series of the effective temperature of the first replica and of all replicas at once are also given. It can be seen that the first replica visited all temperatures several times in the trajectory, proving that the 200-ns-long trajectory is enough to sample all temperatures. The average exchange probabilities were around 0.4, which is a sign that the temperature range was not too wide for the number of replicas used.

A secondary structure analysis for the first replica can be found in Fig. (3.2, c), it reveals that there is a stable α-helix between residues E429 and N436, in agreement with the experiments but spanning two helical turns, whereas the experimental NMR structures only spanned one. The persistence of this helix in the simulations is remarkable as the trajectory contains many high-energy structures that could lead to other local minima lowering the total amount of helix obtained.

As previously stated in the introduction of this chapter, simulating the recognition process of the peptide by CaM is beyond the scope of this work, properly sampling the total conformational space of the two proteins would be extremely challenging even

with metadynamics or other enhanced sampling techniques. Therefore, we skipped that step and assumed that what happens is that the C-lobe of CaM recognizes the fragment containing the pre-helix and binds to it. However, we wanted to test whether a single alpha-helical turn was enough for CaM binding to induce the rest of the helix or whether a greater helical amount was needed. For this reason, we ran REST2 simulations in the presence of the C-lobe of CaM using both the first NMR model (PDB ID: 1KKD) with a single turn and a snapshot from the end of our CaM-free simulations, that had two turns. In the following, we will refer to these two systems as "one turn" and "two turns", respectively. To build the systems we considered that the C lobe of CaM would recognize the fragment in the same place as it appears in the final structure, so we fitted the peptide to the experimental structure (PDB ID: 3SJQ [90]), as depicted in Fig. (3.3, A). As in the previous set of simulations, we solvated a dodecahedral box with TIP3P water molecules [62], assuring that the minimun distance between protein and walls was 1.5 nm. K^+ and Cl^- ions were also added to achieve a physiological concentration of KCl of 0.15 M. The effective temperature range for the REST2 simulations was the same as before, namely 298 K to 373 K, in this case however, CaM C-lobe and the solvent will constitute the "cold" subsystem, whereas the SK2 CaMBD peptide and ions will be the "hot" one, note that the number of replicas needed is related to the extension of the hot subsystem and its interactions with the cold one, which will be similar to the previous simulations in the absence of CaM C-lobe. The counterpart is that the conformational space of CaM-Clobe will be substantially less sampled, so meaningful conformational changes obtained will only refer to the fragment of SK2 CaMBD.

For this set of simulations, a combination of the CHARMM36 force field [21] with GROMACS 2021.3 [23] patched with PLUMED 2.7.2 [87] was used. The workflow followed is the same as for the set of simulations related to Fig. (3.2, C). Ther is also a materials and methods page for these simulations, in which input files for each step can be found together with the time series of the effective temperatures for the first replica ('one turn', 'two turns") and for all replicas ('one turn', 'two turns") [91]. It can be seen that in both cases the first replica properly explores all Hamiltonians so that both high and low energy states have been sampled. It is noteworthy that in this case the exchange probability has increased from around 0.4 in the simulation of the CaM SK2 in absence of CaM to above 0.75 both in 'one turn' and 'two turns" cases; the difference lies in the interfaces between the "hot" and "cold" subsystems, indicating that the interface with water causes more fluctuations in energy than the interface with CaM C-lobe plus water, that subsequently impact on the exchange probability.

The results are displayed in Fig. (3.3, B-D), where we compare the secondary structure of the channel in the absence of CaM Fig. (3.3, B) with the same peptide starting in the "one turn" Fig. (3.3, C) and in the "two turns" configuration Fig. (3.3, D); the last two are in presence of CaM C-lobe, there it can be seen that the rest of the secondary structure can be induced in presence of CaM C lobe for the "two turns" run, in contrast to what happened in the absence of CaM, in which alpha-helical secondary structure in the segment N421-R428 segment seemed to be banned. These results differ from the ones obtained in the "one turn" simulation, where the appearance of helical content is not banned as it appears in different locations throughout the simulation but there is no longer a stable helix along the E429-T437 segment and there are not any frames in which the α-helical content spans the whole SK2 CaMBD, this does not mean that the induction of the whole helix is completely banned, but since high-energy states have been sampled it suggests that it is highly unlikely, making this state less productive in terms of folding.

To elucidate the mechanisms that are trapping the SK2 segment in this unfolded state we computed for each amino acid of the SK2 CaMBD the RMSF and the solvent-accessible surface area (SASA) Fig. (3.3, bottom panel), being the latter the external area that is not buried by other solute atoms. The SASA was calculated because it is known that CaM creates a hydrophobic pocket around its targets, so changes in SASA and RMSF will be indicators of conformational changes. It stands out that W431 exhibits higher mobility Fig. (3.3, E) and less solvent exposure Fig. (3.3, F) in the "one turn" simulation when compared with the "two turns" one. A visual inspection of the trajectories proves that the SASA reduction of the "one turn" simulation is due to the fact that W431 is facing the CaM C-lobe and the RMSF increase is due to it not having secondary structure, in contrast to the "two turns" case, in which W431 is facing the solvent and belongs to an α-helix. Therefore the lack of secondary structure enhances the mobility of the tryptophan and it is kept in the hydrophobic pocket of CaM lowering its SASA. Besides, it coincides that some CaM C-lobe residues that interact with W431 in the "one turn" simulation (this is, when facing CaM C-lobe) interact with L427 in the "two turns" one , namely F92, M109, L112 and M145, suggesting that in the "one turn" simulation, W431 has competed with L427's binding pocket leading to a less productive conformation.

The previously discussed competition for the CaM C-lobe residues takes place in the in the "one turn" simulation, in which W431 is facing CaM. We wanted to see

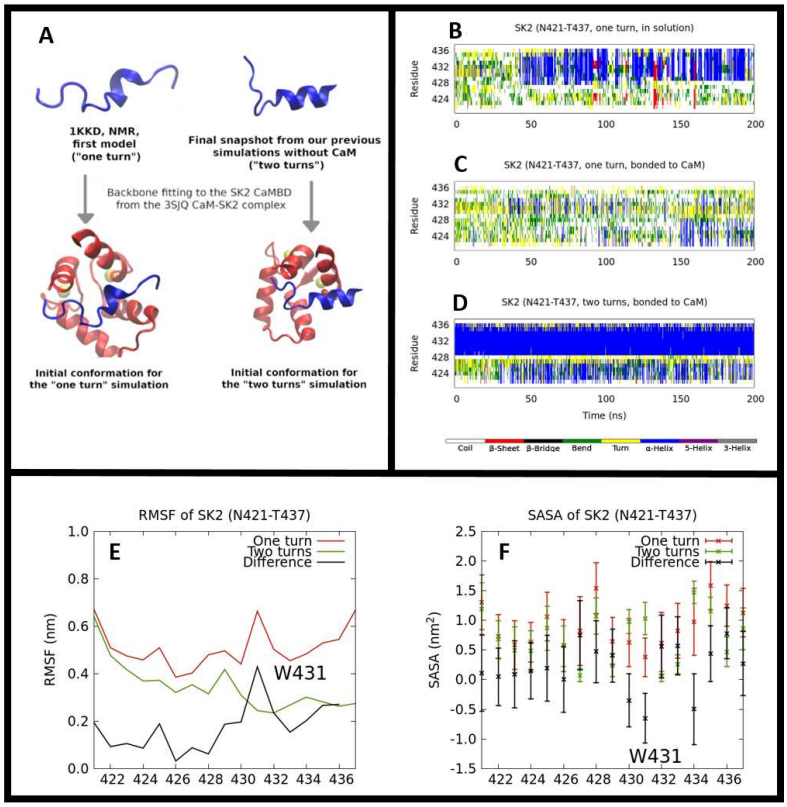


Figure 3.3: **A**, workflow followed and depiction of the constructs composed by the SK2 CaMBD and CaM C lobe complex, for the "one turn" and "two turns" conformations. **B-D**, secondary structure of the first replica according to DSSP definition [86] of the SK2 immerse in water in the absence of CaM (**B**) compared with the "one turns" conformation of the SK2 CaMBD fragment bound to CaM (**C**) and the "two turns" conformation of SK2 CaMBD fragment bound to CaM (**D**). **E-F**, analysis of the "one turn" (red), "two turns" (green) simulations and their difference (black) for the amino acids belonging to the SK2 CaMBD, highlighting the amino acid W341. The computed magnitudes are RMSF (**E**) and SASA (**F**). Adapted from [71]

if, by preventing this competition, the rest of the helix could be induced, so we rotated W431 by 180 degrees, forcing it to point outwards, this is, towards the solvent Fig. (3.4, A-top). The fact that we obtained a less productive state with W431 facing CaM was shocking as experimental conformations of the CaM-SK2 complex show that W431 indeed points to CaM (PDB: 3SJQ, [90]), so we also rotated the "two turns" configuration to point inwards to CaM to see if the α-helix induction holds also for this configuration Fig. (3.4, A-bottom).

REST2 simulations were carried out for these two new "swapped" configurations. The building of the systems, software and workflow is the same as for the previous set of simulations. As usual, supplementary information can be found at its materials and methods github page [92], namely input files of all steps of the workflow with snapshots and the time series of the effective temperature for the first replica ("one turn", "two turns") and for all replicas at once ("one turn", "two turns").

The secondary structure analysis for theses simulations is presented in Fig. (3.4, C-D), with that of the channel segment immersed in water in absence of CaM Fig. (3.4, B). Differently to the previous case, the "one turn" simulation produces the whole helix when W431's side chain is pointing outwards, furthermore, the helical content starts forming around this tryptophan Fig. (3.4, C). For the "two turns" simulation with W431 pointing inwards Fig. (3.4, D), the secondary structure obtained is similar to the previous case in which the side chain was pointing outward; the helical content of E429-N436 segment is maintained and the rest of the helix can be induced. In this second set of simulations, the RMSF and SASA were also computed Fig. (3.4, bottom panel), indicating that the tryptophan has now similar mobility when pointing inwards and outwards Fig. (3.4, E) and the SASA difference is now not conclusive due to its large error bars Fig. (3.4, F).

Putting the results together, we have seen that W431 pointing inwards starting from the "two turns" configuration becomes productive. This is translated into a descent in the mobility of the tryptophan from $\approx$ 0.7 nm in the "one turn, inward" Fig. (3.3, E) to $\approx 0.3\,nm$ in the "two turns, inward" Fig. (3.4, E), since the tryptophan is constrained into a helical conformation. It is also noteworthy that when looking at the two trajectories that had W431 facing CaM, there is an increase of SASA from $0.4 \pm 0.3\,nm^2$ in the "one turn, inward" system Fig. (3.3, F) to $0.9 \pm 0.5\,nm^2$ in the "two turns, inward" system Fig. (3.4, F), meaning that W341 was better buried in the "one turn" simulation. This SASA difference can be related to the contacts stablished with CaM in both situations, in the "one turn, inward" system, we had that the CaM residues A88, V91, F92, M109, L112, F141, and M145 were in contact with W341, but in the "two turns, inward" system, these contacts are reduced to M145 alone. This time, it is the SK2 residue V426 (instead of L427) the one that now makes contacts with CaM in the frames that the full helix is induced, being these E120, E123, M124, and E127.

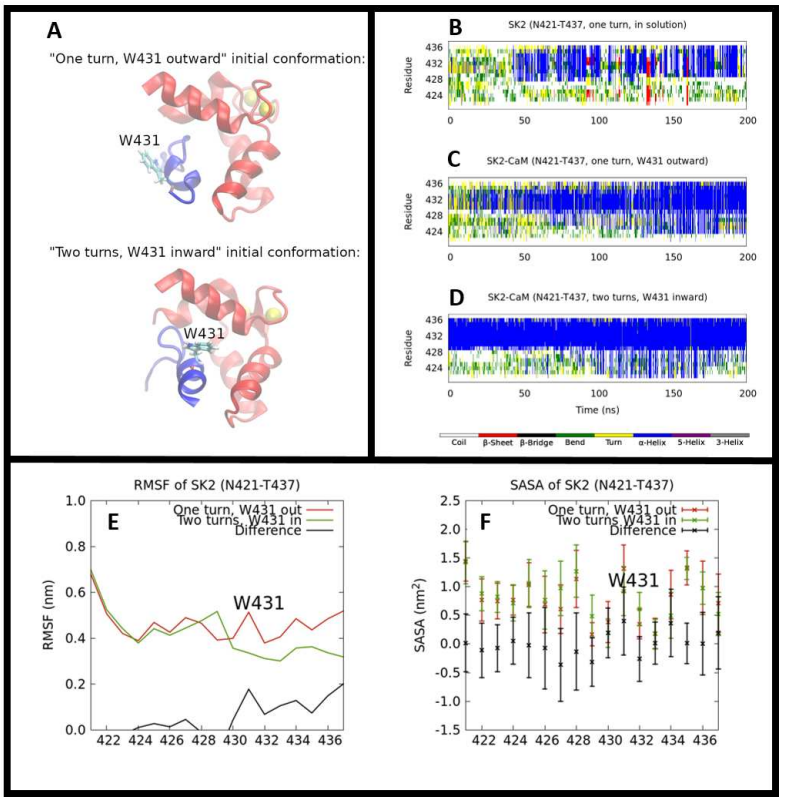


Figure 3.4: **A**, starting structures of the second set of simulations, in which the SK2 CaMBD was rotated by 180°. **B-D**, secondary structure of the first replica according to DSSP definition of the SK2 immersed in water in the absence of CaM (**B**) compared with the "one turn" conformation of the SK2 CaMBD fragment bound to CaM with the W341 pointing outwards, this is, towards the solvent (**C**) and the "two turns" conformation of SK2 CaMBD fragment bound to CaM with the W341 pointing inwards, this is, towards CaM C-lobe (**D**). **E-F**, analysis of the "one turn" (red), "two turns" (green) simulations and their difference (black) for the amino acids belonging to the SK2 CaMBD fragment, highlighting the amino acid W341. The computed magnitudes are RMSF (**E**) and SASA (**F**). Adapted from [71]

Altogether, it has to be highlighted that the previous REST2 simulations in the presence of CaM C-lobe did not achieve its final purpose, this is, efficiently sampling the conformational space, since simulations starting from different conformations raised different results. More precisely, changes in the orientation of the tryptophan were not observed, even though it seems to be important in the folding process. It is likely that the energy barrier that prevents this spin to happen is bigger than the

thermal energy of the highest effective temperature that we could afford (373 K). To overcome this limitation and confirm the presence of an energy barrier separating those states we needed a different approach. Metadynamics simulations turn out to be appropriate.

The systems used where the previously analyzed "one turn" and "two turns", which were already minimized and thermalized. As before, simulations were ran with the CHARMM36 force field [21] in GROMACS 2021.3 [23] patched with PLUMED 2.7.2 [87]. For defining the collective variable for the addition of the Gaussians, the angle formed by the center of mass of CaM C-lobe, C_α carbon of W431 and the center of mass of the hexagonal aromatic ring of W431 was selected, the collective variable is depicted in Fig. (3.5, A). Gaussian functions were added to the Hamiltonian every picosecond, each centered around the value of the angle at that simulated time and with a height of $2.0\,kJ/mol$ and a width of 0.5 rad. We observed that sometimes the addition of these Gaussians provoked the dissociation of the CaM C-lobe and the channel fragment, so we had to add a funnel-like potential to prevent it. 2500-ns-long, well-tempered metadynamics simulations in the NPT ensemble showed two differentiated minima, one near $0\,rad$ of the previously described angle, this is pointing towards CaM, and another near $\pi\,rad$, pointing towards the solvent, for both the "one turn" and the "two turns" starting conformations, even though in the latter the minimum towards CaM is more spread Fig. (3.5, A). While these simulations are not fully converged (see the supplementary time series and convergence picture of the collective variable), it is noteworthy that, in both cases, the orientation of the tryptophan towards CaM seems to be more stable.

Further information of these simulations can be found in its materials and methods github page [93], in particular the input files needed by gromacs to perform the MD and also the input for the PLUMED plugin that contains the metadynamics parameters, together with the time series and convergence of the collective variable for both simulations.

With the information of the stability upon orientation angle and the secondary structure analysis we can sketch the recognition mechanism of the SK2 channel by CaM C-lobe Fig. (3.5). Conformational pool of the SK2 CaMBD fragment is mainly formed by the "one turn", "two turns" or coiled like conformations, according to the NMR experimental results Fig. (3.2, A) and our REST2 simulations in the absence of CaM Fig. (3.2, C). From these conformations CaM would select the productive confor-

mations for SK2 CaMBD, namely both "two turns" conformations and the "one turn" that has the tryptophan pointing towards the solvent. As revealed by the REST2 simulations, the rest of the α-helix would be induced afterwards, and in the case that W341 is facing the solvent, it would have to rotate to point CaM, since experimental structures have this tryptophan facing the CaM C-lobe [90].

Gathering the results from the simulations, it seems that the three most important channel residues for inducing the helix are W431, V426, and L427; in fact, X-ray studies have determined L427 and W431 to be the residues that interact the most with the hydrophobic pocket of CaM C-lobe [90] and NMR experiments showed that mutating V426 and L427 to glycines prevented the appearance of the pre-helix of the SK2 CaMBD [2]. In our simulations, the role of these three residues can be sketched as it seems that V426 and L427 need to make contacts with the hydrophobic pocket of CaM, but there can be situations in which W431 can compete for these contacts making the state less productive.

Finally, with the purpose of generalizing the interesting results obtained for the SK2 channel to other CaM targets, REST2 simulations of several CaMBD fragments (see Table 3.1) immersed in water were carried out, following the same methodology as in the simulation referring to Fig. (3.2, C) [37]. The input parameters of each step of the workflow for each simulation can be found in its materials and methods page [94]. The first step to generalize our results was to include the rest of the KCNN family members, this is, the KCNN1,KCNN3 and KCNN4 channels. A further generalization was to include some constructs that contain the so-called IQ-like binding motif, such as KCNQ1, KCNQ2, MYH7, CAC1C, INVS or IQGA2. Further information of the analyzed sequences can be found in Table 3.1. In order to assure that all effective temperatures were sampled in all systems, the time series of the temperature indexes of the first replicas are provided[2], also exchange probabilities are provided[3]. The secondary structure analysis of these targets aredisplayed in Fig. (3.6, A-H). It is interesting that residues V426, L427 and W431 are conserved through the whole KCNN family and the appearance of a robust pre-helix is also maintained, whereas those structures that only have W431 such as KCNQ1 and KCNQ2 or do not have any of the three residues, like MYH7, CAC1C, INVS, or IQGA2, show some helical content

[2]Follow the links for exchange time series of KCNN1, KCNN4 KCNQ1, KCNQ2, MYH7, CAC1C, INVS or IQGA2

[3]Follow the links for exchange probabilities of KCNN1, KCNN4 KCNQ1, KCNQ2, MYH7, CAC1C, INVS or IQGA2

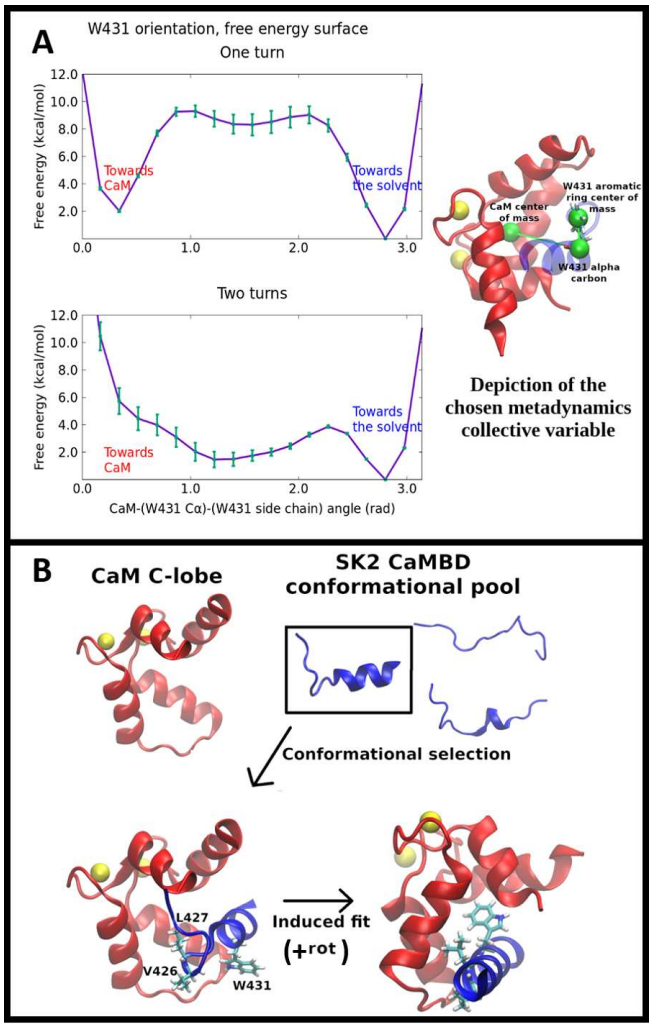


Figure 3.5: **A**, free energy profile projected on the angle between the CaM C-lobe, the C_α of W341 and its aromatic ring, in rad, computed by means of metadynamics simulations for both the "one turn" and "two turns" conformations, along of 2.5 μs long simulations. The error bars represent the standard deviation of the free energies along the last 200 ns. **B**, depiction of the mechanism of the SK2 CaMBD fragment recognition by CaM C-lobe that arises from our simulations, which is a combination of the selection of the conformations of SK2 CaMBD that have two helical turns and the induction of the rest of the helical content together with a rotation if the selected conformation has W431 pointing to the solvent. Adapted from [71]

but it is not as persistent as in the KCNN family. These results suggest that the previously described mechanism can be generalized to the entire KCNN family since

they exhibit similar behaviour (Fig. 3.2, C; Fig. 3.6 A-B), and that could also be present in the IQ motifs, since the ones analyzed here are also able to form pre-helices when immersed in water in the absence of CaM Fig. (3.6, C-H), even though they are not as persistent as those of the KCNN family.

Entry name	ID	Sequence	Nº residues
KCNN1_HUMAN	Q92952	NAAANVLRETWLIYKHT	17
KCNN2_HUMAN	Q9H2S1	NAAANVLRETWLIYKNT	17
KCNN3_HUMAN	Q9UGI6	NAAANVLRETWLIYKHT	17
KCNN4_HUMAN	O15554	ESAARVLQEAWMFYKHT	17
KCNQ1_HUMAN	P51787	PAAASLIQTAWRCYAAE	17
KCNQ2_HUMAN	O43526	NPAAGLIQSAWRFYATN	17
MYH7_HUMAN	P12883	SRIITRIQAQSRGVLAR	17
CAC1C_HUMAN	O13936	FYATFLIQEYFRKFKKR	17
INVS_HUMAN	Q9Y283	DIAAFKIQAVYKGYKVR	17
IQGA2_HUMAN	Q13576	EENVVKIQAFWKGYKOR	17

Table 3.1: Simulated constructs, referred to Fig. (3.6). Data provided are the UniProt entry name, UniProt entry ID, simulated sequence and number of residues of each construct, information extracted from the UniProt database (https://www.uniprot.org/), Uniprot entry IDs are linked to their data base page. Residues at positions 426, 427 and 431 are highlighted for those sequences including V, L and W, respectively. Note that KCNN1 and KCNN3 have the same sequence, so only one (KCNN1) was simulated.

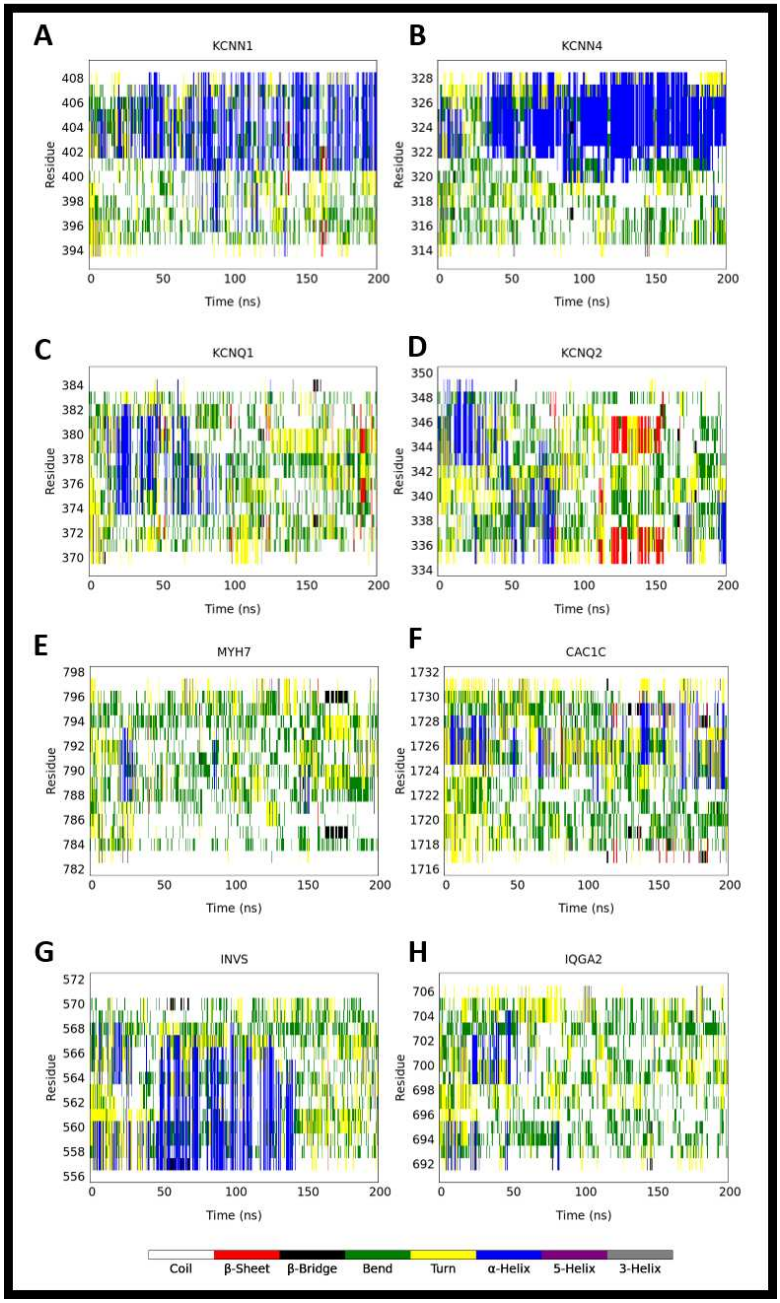


Figure 3.6: **A-H**, secondary structure according to the DSSP definition [86] of some CaM targets immersed in water and in the absence of CaM. Adapted from [71]

3.3 Conclusions

In this chapter, a computational effort has been done in order to gain insight into how CaM recognizes its targets, focusing on the channel KCNN2. Even though we have assumed the starting conformation of the system composed by the SK2 CaMBD fragment bound to the CaM C-lobe, enhanced sampling techniques had to be used in order to observe α-helix formation; for that purpose, HREX REST2 simulations were carried out. In the meanwhile, we realized that the orientation of W431 was crucial for helix induction in the "one turn" simulations. However, angle transitions were not observed throughout the simulations and we had to estimate the energy barrier through metadynamics simulations.

Our results indicate that the recognition process takes place following a mixture of the two established theoretical frameworks: conformational selection and induced fit. There is both experimental and computational evidence that the SK2 CaMBD fragment exhibits some helical content when immersed in water, Fig. (3.2, A) [2] and Fig. (3.2, C), respectively. Suggesting that these helical turns are needed for the subsequent recognition (conformational selection), but the apparent inability of forming more helical content would imply that CaM is needed to achieve the full secondary structure observed in experimental structures [76, 90]. In fact, the mechanism by which this induction takes place has been observed in our HREX REST2 simulations (Figs. 3.3, 3.4), also arising conformations that are not productive in terms of α-helix formation, like the "one turn" pointing towards CaM. Aiming to generalize our results, other CaM targets immersed in water were simulated to check if they were able to form some helical content in the absence of CaM Fig. (3.6). From the results it was clear that the other members of the KCNN family have a similar behavior, and presumably their α-helix will be induced in a similar manner. In contrast, other CaM targets that contain an IQ-like motif exhibited a less persistent pre-helix, but it has to be emphasized that helical content was observed in all targets, making us think that the previously described process for target recognition would hold for them, and the difference in pre-helix persistence could be related to the time extent of the recognition process.

Chapter 4

The W344R mutation revisited

4.1 Introduction

In this chapter we will revisit the W344R mutation of the $K_V7.2$ channel, by applying the methodology introduced in the previous chapter, this is, we will try to elucidate the effect of the mutation in secondary structure formation through REST2 Hamiltonian replica exchange simulations. The biological significance and experimental evidence behind the W344R mutation was extensively reviewed in chapter 2, where we concluded that this mutation is quite peculiar as it has different effects on CaM binding depending on the nature of the experiment. More precisely, CaM binding is enhanced in *in vitro* experiments but disrupted in *in vivo* experiments [1]. In this chapter we will design two sets of simulations that resemble the conditions of both types of experiments, with the aim of observing differences in secondary α-helix that could be relevant for CaM recognition. The motivation of these calculations was to improve the two-state-conformations hypothesis formulated in chapter 2, whose limitations were commented in the conclusions section of that chapter.

4.2 Simulations in water

In the previous chapter, we validated a protocol to observe secondary structure formation of small fragments in water, based in the REST2 enhance sampling technique [80].

In the following set of simulations we have used the same methodology, with the aim of reproducing the *in vitro* conditions of a small mutated fragment of the $K_V7.2$ channel near the mutation site, whose sequence is N334-PAAGLIQSA**R**RFYAT-N350. Afterwards, the results will be compared with the wild type ones, N334-PAAGLIQSA**W**RFYAT-N350, where the output of the latter was already given in Fig. (3.6, D). The building of this fragment in aqueous solution was carried out by introducing the mutation to the wild type system, whose building was described in detail in the previous chapter, and is also given in the supplementary materials github page, together with the details of the protocol and input files of the simulations. As usual, the simulations were carried out using GROMACS 2016.4 [16] with the CHARMM36 force field [21].

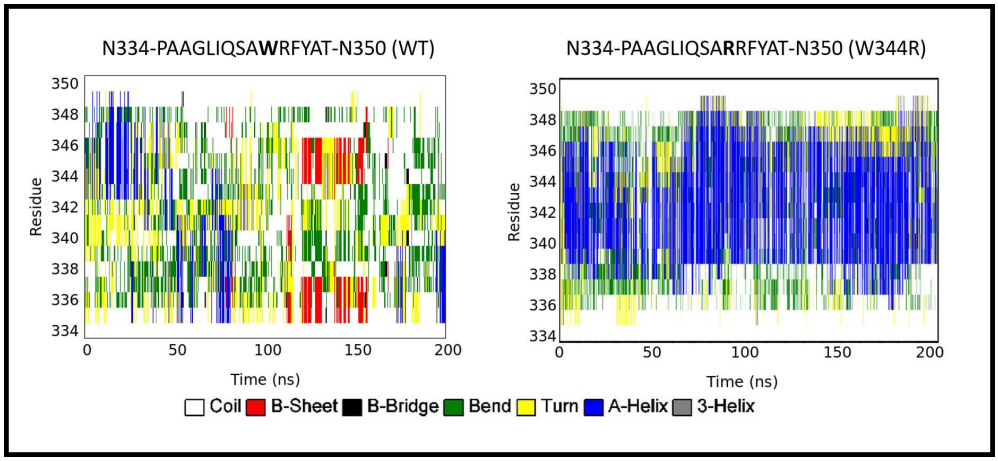


Figure 4.1: Secondary structure of the first replica of the REST2 simulations throughout the simulated time, according to the DSSP definition [86]: wild type fragment (left) compared with the W344R mutation (right).

From the simulations, the secondary structure of first replica was computed using the DSSP definition [86] and was plotted throughout the simualted time in Fig. (4.1). It is clear that the W344R mutation enhances α-helix formation when the fragment is immersed in water. Exchange probabilities of the mutant were larger than those of the WT, with a probability increase of about $\sim 20\%$, but both simulations were able to explore the temperature space, as can be seen for the first replica of WT and W344R and for all at once of WT and W344R. We hypothesize that this situation would take place in *in vitro* experiments; the W344R mutation would bias the CaMBD towards a folded state that would be more easily recognized by CaM, increasing the experimental CaM binding observed in *in vitro* experiments [35].

4.3 Simulations in the ribosomal tunnel

The second part of this chapter consists in trying to simulate the cotranslational effect introduced by the W344R mutation. For experimental reproducibility, we have created a construct with the SecM arresting peptide linked to a small fragment of the $K_V7.2$ channel, which will conform the nascent chain, so that we are simulating a static system in which part of the S6 helix of the transmembrane domain and of the hA helix of the CRD are arrested in the ribosomal tunnel [42]. The fragment of the channel was chosen so that the 344 position was in the vestibule.

Before analyzing the results obtained, it is interesting to detail how these simulations were set. The starting structure was one cryo-EM structure of the ribosome of *escherichia coli (E. Coli)* that includes a part of the SecM arresting peptide in the ribosomal tunnel (PDB ID: 3JVB) [95]. The $K_V7.2$ channel sequence was added to this SecM peptide by using the "build" command in Pymol 2.4.1 [96], until we obtained the situation of Fig. (4.2, A). However, in order to obtain a final structure with an affordable number of atoms for subsequent HREX simulations, the ribosome tunnel needed to be cropped. For doing so, the reduction was performed in the three dimensions; on the one hand, only amino acids or nucleic acids with at least one atom less than 1.2 nm away of the nascent chain were kept. This length was slightly increased to 2.0 nm for the amino acids and nucleic acids conforming the exit of the tunnel, trying to conserve as many interactions as possible between the part of the nascent chain outside the tunnel and the ribosome. On the other hand, the highlighted proline in Fig. (4.2, A) was selected as the end of the tunnel, meaning that the rest of the SecM after this proline was removed from the system Fig. (4.2, A: sequence in lighter purple) and it was taken as reference to crop the ribosome in the nascent chain direction. When carrying out this reduction of the ribosome, amino acids or nucleic acids were not broken, meaning that if one atom belonging to any amino or nucleic acid satisfies one of the previous conditions the whole amino or nucleic acid is kept in the system. However, when chains were broken, meaning that we broke peptide bonds in proteins or phosphodiester bonds in RNA, patches were added to avoid the introduction of extra charges in the system. Side and bottom views of the reduced ribosome can be seen in Fig. (4.2, A and B) respectively, where the kept nucleic acids are depicted in brown surface and the kept amino acids in orange, and the position of W344 is highlighted in the nascent chain. The reduction of the ribosome was performed with VMD [33]. Next, the system was placed into a dodecahedral box with

a minimum distance between its sides and any nucleic acid or protein atoms of 1.0 nm, which was filled with of TIP3P water molecules [62]. A 0.15 M concentration of KCl was added to mimic neuron physiology, and some additional potassium ions were added to neutralize the total charge of the system. This resulted in the wild type structure depicted in Fig. (4.2, C). The W344R mutation was introduced with Pymol 2.4.1's "mutagenesis" command [96] and is shown in Fig. (4.2, D). Finally, GROMACS topologies of the systems were generated with CHARMMGUI [97].

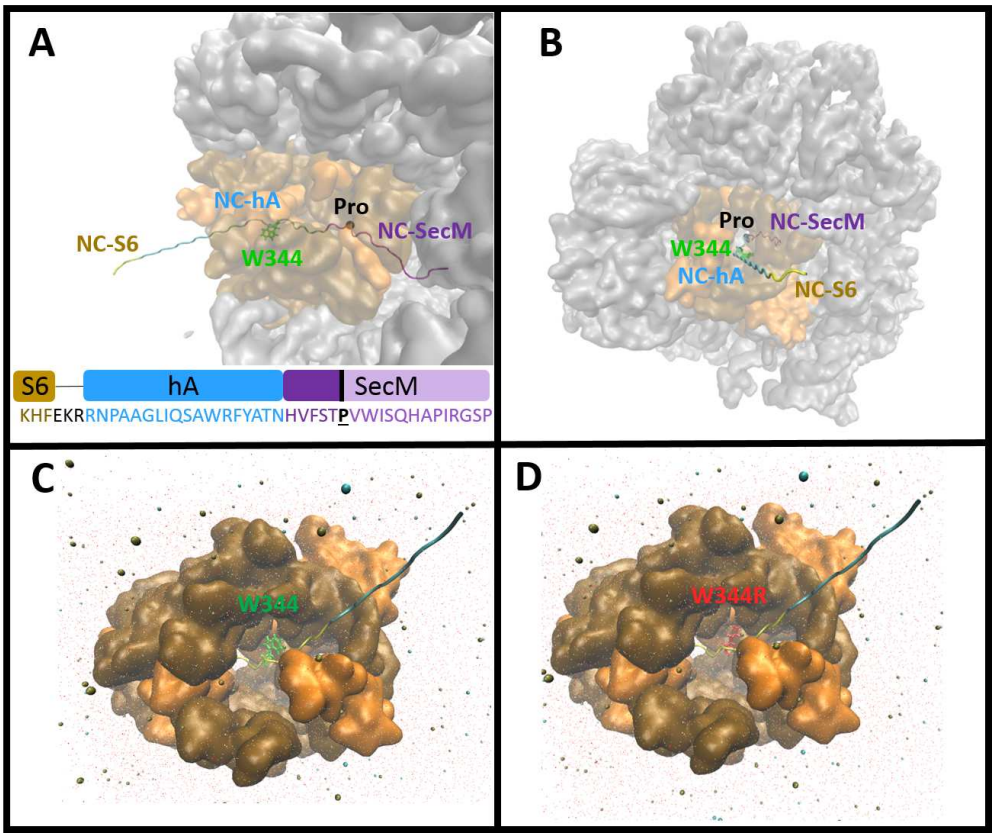


Figure 4.2: **A** - **B**, side view and top view of the system formed by the ribosome (gray area) and the nascent chain. Orange and brown areas are respectively proteins and nucleic acids forming the tunnel kept in the simulations. The nascent chain sequence is provided in **A** and spans part of the S6 transmembrane helix and of the hA in the $K_V7.2$ channel linked to the SecM arresting peptide, where W344 is also showed. The highlighted proline of SecM was used for cropping the ribosome. Starting conformations of the simulated systems: WT (**C**) and W344R mutation (**D**), the "hot" subsystem is colored in yellow.

For these systems, REST2 simulations were carried out to explore the conformational space of the nascent chains of wild type and mutant with special interest in secondary structure formation. For that purpose, 15 replicas were set spanning a range of effective temperatures of 300K to 450K, where the temperature of replica i (i from 0 to 14) is given by $T_i = T_0(\frac{T_{14}}{T_0})^{i/14}$. In this case, the "hot" subsystem of the Hamiltonian is the same fragment that was simulated in water, and it is depicted in yellow in Fig. (4.2, C and D), whereas the "cold" subsystem is the rest. For equilibrating the systems the following protocol was employed, it is also described in the github supplementary information page. The first step was to minimize the system without any constraints, so that the atoms of the ribosome, nascent chain and solvent could find a local minima that avoids atomic clashes. After the minimization, all simulations were performed with physical restraints on the ribosomal atoms, since we have cropped it, and on the first proline of the nascent chain, which belongs to the SecM, so that the nascent chain does not detach from the ribosome throughout the simulation. Then, to equilibrate the solvent and adjust the simulation box, a 100-ps equilibration in the NVT ensemble where followed by another 200-ps equilibration in the NPT ensemble, in both cases physical restraints on the nascent chain and ribosome were imposed, with constants of 10000 $kJmol^{-1}nm^{-1}$. To check the effects of the restrictions, RMSD of WT and W344R during thermalizations are provided as supplementary material. As usual, temperatures were kept at 298 K using a velocity-rescaling thermostat [88] and pressure at 1 bar using the Parrinello-Rahman barostat [89] in the NPT stage, all these steps were performed with the base Hamiltonian to provide the starting conformation of all replicas in the REST2 simulations. Finally, three independent 300-ns- long REST2 simulations were carried out, so that the total collected time with the base Hamiltonian was 900 ns, all in the NPT ensemble as before but without restrictions on the nascent chain, aside from its first proline. In fact, the effect of removing the position restraints can be checked in the RMSD[1] and RMSF[2] plots, where it can be seen that most part of the contribution to the RMSF comes from the amino acids K327 and H328, which are outside of the tunnel. The exchanges probabilities were analyzed, these probabilities were around $\sim 60\%$ for both WT and the W344R simulations, even though the temperature interval had been increased in this set of simulations compared to previous ones. This was translated into a proper

[1] Follow the links for RMSD of WT-replica1, WT-replica2, WT-replica3, W344R-replica1, W344R-replica2, W344R-replica3

[2] Follow the links for RMSD of WT-replica1, WT-replica2, WT-replica3, W344R-replica1, W344R-replica2, W344R-replica3

exploration of all effective temperatures by the first replica[3], except for the third run of the mutant simulations. However, it was able to reach to the highest temperature twice. As usual, exchanges of all replicas are also given[4].

Throughout the simulations, cotranslational helix formation was observed for both the wild type and the mutant constructs. This cotranslational helix formation was computed with DSSP [86] and it is characterized in Fig. (4.3, A), where the percentage of frames in which helical content appears for wild type and mutant is plotted, for each amino acid of the REST2 "hot" subsystem.The helical content has been categorized into α-helix and 3_{10} helix, where the difference between both secondary structures elements is the number of amino acids forming a helical loop. It has to be underlined that the total helical content of WT almost doubles that of W344R, even though the number of frames in which helical content was observed is low, about $\sim 3\%$, which is a clear indicator that the ribosomal tunnel has a destabilizing effect on helix formation. This effect has been reviewed in [98]. It happens for several proteins, where the cause seems to be the net negative charge of the ribosomal tunnel, the goal of this effect would be to avoid intermediate folding states that would avoid the formation of the fully folded state.

It is also interesting to point out the effect of the mutation at the location of the helices corresponding to peaks W344-1 and W344R-1 in Fig. (4.3, A). The mutated arginine seems to displace the location of the helix, since it was not found to have secondary structure throughout the simulations. In fact, the representative conformation of the W344R-1 peak show the helical content stops forming after the arginine, in contrast to what happens in W344-1, where the tryptophan is part of the helix and as a result it is located deeper in the ribosomal tunnel, both conformations are showed in Fig. (4.3, A, right structures) .

To a lesser extent, there are a few conformations (representing around 0.25% of the total simulated time), named W344-2 and W344R-2, in which helical content is found along the P335-L339, residues which are located closer to the exit tunnel, and in these conformations there are no significant differences between wild type and mutant, which could be an indicator that the effect of the W344R mutation is mainly local.

With the aim of elucidating the mechanistic differences observed in the secondary

[3]Follow the links for exchanges of the first replica of: WT-replica1, WT-replica2, WT-replica3, W344R-replica1, W344R-replica2, W344R-replica3

[4]Follow the links for exchanges of all replicas of: WT-replica1, WT-replica2, WT-replica3, W344R-replica1, W344R-replica2, W344R-replica3

structure formation, the contact probabilities between both nascent chains and the ribosomal tunnel depicted as a circos plot in Fig. (4.3, B:left and right), respectively

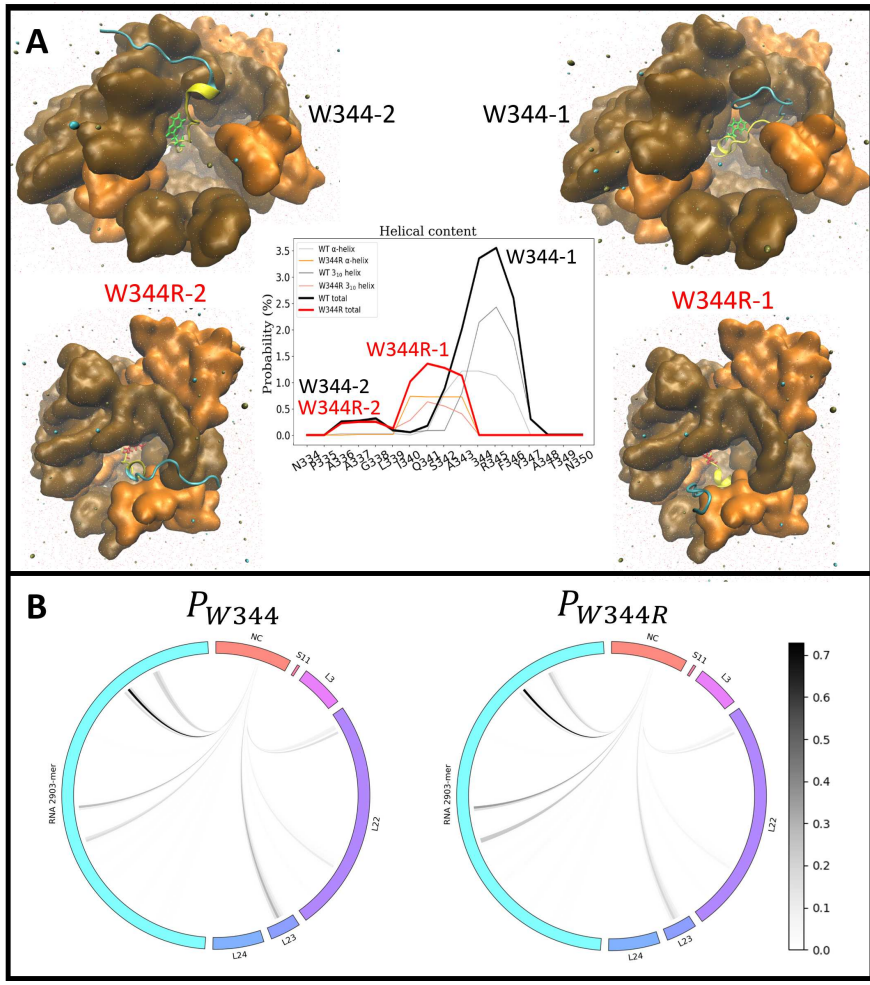


Figure 4.3: Results of the REST2 cotransaltional simulations of wild type KCNQ fragment and the mutated W344R fragment. **A**, Helical content probability for the wild type (black) and the W344R mutant (red), for all amino acids of the "hot" subsystem of REST2 simulations. All peaks are labeled and one of their most representative conformations is depicted, following the color code introduced in Fig. (4.2). Circos plot of the contact probability of W344 (left) and W344R (right) with all ribosomal amino acids and RNA. In the outer circle the whole simulated sequence is depicted, with the nascent chain (NC) represented in red, then the ribosomal proteins S11-L24 and finally the RNA sequence is represented in cyan, the black lines connect position 344 with the contacted residue with the color depending on the contact probability.

[99].

From a visual inspection we can conclude that there are no major differences in the contacts of the mutant and wild type with the ribosomal tunnel.

Finally, a comment on the convergence of these simulations should be made, we conducted three replicas for both the wild type and the mutant of our concerns about the reliability of the results due to the small helical content observed. To assess this issue, we plotted the helical content obtained through different simulation times to see if the final obtained results seem to be converged, this picture is in the supplementary information page and clearly shows that the results of wild type and W344R are converging towards very different values: as more time is simulated the percentage of frames with helical content for the mutant are constantly diminishing, in contrast, those for the wild type are constantly increasing. Therefore, with this simulated time, the difference between the two systems seems clear, even though the calculated percentages have to be considered as approximated.

4.4 Conclusions

In this chapter, we have assessed the biological problem of the W344R mutation of the $K_V7.2$ ion channel from a different perspective, based on a secondary structure analysis instead of the two conformations described in the previous chapter, whose limitations were also commented in its conclusions section. We have reproduced the conditions of the *in vitro* experiments by simulating a fragment of the channel in aqueous solution and the *in vivo* ones by simulating the fragment inside the ribosomal tunnel. The main hypothesis behind this approach, based on the experimental evidence and on our computational analysis presented in the previous chapter, is that some helical content is needed for subsequent CaM recognition and folding of the channel. Within this framework, the W344R mutation has been observed to significantly enhance the α-helix formation for the simulated fragment in aqueous solution. Which we attribute to the better CaM binding that this mutation has showed in *in vitro* experiments. The simulated fragment is small but we believe that it can be representative, as this fragment is critical for CaM binding. Moreover, the computational cost of REST2 severely restricted the maximun affordable size of the simulated fragments.

Regarding the fragments inside the ribosomal tunnel, we knew from the previous chapter that we were going to achieve nice exchange probabilities, as the "hot"

subsystem was similar to the ones of the previous chapter, so we increased the temperature range to further enhance the sampling. However, the obtained helical content was drastically reduced when compared with the simulations in aqueous solution, but some differences were found between WT and W344R. The WT fragment exhibited a greater helical content probability (even though the total the probabilities were low) and the mutated arginine shifted it in the sequence. The smaller amount of helical content could be associated to an impaired CaM recognition and could explain the experimental results obtained *in vivo* in which CaM binding was substantially reduced and led to a nonfunctional channel. However, to further test this hypothesis, more simulations should be carried out at different points in the translation of the channel to check if the small differences observed in helical content become larger at more advanced stages of the synthesis. Additionally, other mutations could be simulated immersed in aqueous solution, to see if there is a correlation between folding and CaM binding, for example the I340E mutation was found to be very deleterious to CaM binding and channel functioning, so it could be used as a negative control [35].

To combine the results obtained with those of chapter 2, we would need a helical turn spanning Q341 to W344R, as we defined the angles with the C_α of Q341 and W344R when both were forming a helix. This was not seen in the cotranslational simulations, as the W344R seems to move the location of the α-helix, making the angles inconsistent. Nonetheless, we can keep the two types of interactions observed for the mutated arginine[5], and hypothesize that the helical content is the main feature distinguishing these two channel variants and making W344R interact like the wild type or not interact at all with CaM.

[5]Those were a native-like, strong interaction with CaM, leading to charge compensations and a non-native-like interaction, mainly with water

Chapter 5

MD analysis of activators of

5.1 Introduction

The following chapter will be the last one of this thesis and it is separated from the previous ones, that refer mainly to cotranslational folding and secondary structure adoption. This chapter is a computational effort to try to elucidate the effect of some drugs on the full channel structures of the $K_V7.2$ and SK4 ion channels. Lately, as a consequence of the revolution that the cryo-EM field is experiencing, a wide variety of full ion channel structures have been reported with an outstanding resolution, in particular $K_V7.1$, $K_V7.2$ [13, 34] and small conductance calcium activated SK channel KCNN4 (also known as SK4) [76]. Furthermore, these structures have been obtained in the presence of cofactors, like the PIP_2 lipid, which plays a role in many signaling events, or even drugs, offering the possibility to study their effect from a computational perspective. Recall that cryo-EM structures only contain static information of a stable conformation, meaning that some dynamical information can be obtained through simulations to take the understanding of ion channels one step beyond.

Simultaneously, the increment of accessibility to computational power has made it possible to simulate these full channel structures for periods of time of hundreds of nanoseconds [100], or to build models by homology when only similar structures to the target one were available [58]. In our case, it was possible to acquire the computational power needed thanks to the Spanish high performance computing network (RES), otherwise we would not have been able to perform the simulations.

In the case of the $K_V7.2$ channel, the two avalaible drugs were retigabine (RTG,

PDB ID:7CR7) [34] and ZTZ240 (ZTZ, PDB ID: 7CR4) [34], which have been reported as activators of the channel [34]. Furthermore, they have different effects on the channel, on the one hand, ZTZ increases the maximum current amplitude and enhances the voltage sensitivity so that less voltage differences are needed for gating, whereas RTG only enhances voltage sensitivity. The two drugs presumably alter the channel functioning in a different way. The fact that theri binding pockets lie in different domains support the idea that their mechanisms of actions are different. RTG binds near the pore (see Fig. (5.1, A: cyan surface)), between helices S5, pore helix, S6 (red) and helix S6' of the adjacent monomer (blue). It maintains the tetramerical arrangement of the channels as up to four RTG molecules can be found in the experimental structure (see Fig. (5.1, B)). In contrast, ZTZ is bound to the voltage sensor domain of the channel (Fig. (5.1, C: cyan surface)) and the drug also appears in the four monomers (Fig. (5.1, D)).

To complement the variety of drugs under study, we added riluzole (RIL) to the project. It is known to be the first FDA-approved treatment for amyotrophic lateral sclerosis, and current evidence indicates that this drug works by allosteric activation of KCNN channels[3], as its binding pocket has been experimentally resolved in the SK2 channel lying in the cytosolic domain, and more specifically in the calmodulin binding domain [3]. It was possible then to build a model of the full channel structure of SK4 [76] with RIL in the binding pocket defined by a reduced SK2 structure [3], since the full channel of SK2 channel is not available. As depicted in Fig. (5.2, A: cyan surface), RIL finds its binding pocket between the N-lobe of CaM (yellow) and a small helix belonging to the voltage sensor domain of the channel (grey). Following the observed behaviour of the drugs in the $K_V7.2$ channel, we placed one RIL molecule in each monomer (see Fig. (5.2, B)). In this picture, the bottom view of the channel is shown for a better visualization of the pocket, since RIL is located in the cytosolic domain.

The question that we wanted to address in this project is whether it is possible to characterize the effect of these drugs through simulations. We have to keep in mind that MD simulations have limitations. For example, we are restricted to time scales in the order of hundreds of nanoseconds, this is not enough for such large structures to experience major conformational changes that could result in the experimentally observed effects for the drugs. Moreover, the amount of data obtained from such large simulations can be really cumbersome, as the variety of features that could be ana-

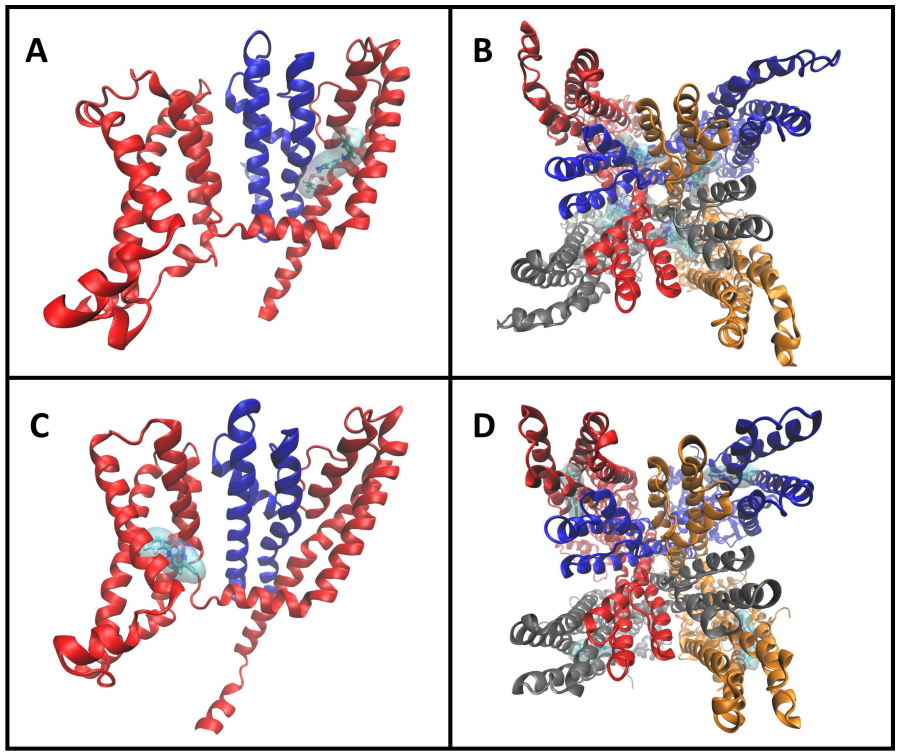


Figure 5.1: Graphical representation of the binding pocket of the drugs RTG and ZTZ in the $K_V7.2$ channel. **A**, insight of the binding pocket of RTG (cyan surface), between helices S5, pore helix, S6 of the voltage sensor domain (red), and the helix S6' of the adjacent monomer (blue). **B**, top view of the channel to show the four RTGs bound to the channel. **C**, insight of the binding pocket of ZTZ, in the voltage sensor domain (red). For consistency, the adjacent monomer is also showed (blue). **D**, top view of the channel bound to ZTZ, with the drug appearing in the four monomers.

lyzed is enormous considering that the drugs have to work allosterically, considering that the binding pockets of the drugs lie far from the pore but they actually influence its opening. In order to deal with these two limitations, a computational graph-based approach called "Network Analysis" [100] was created.

In this chapter, we will describe the methodology followed to build the systems and run the simulations, as some bioinformatic tools were used. Then we will report a characterization of the binding pocket of RIL in the SK4 channel, as there were some dynamical properties that suggest a possible mechanism of action. Finally, the data obtained from performing network analysis will be presented together with a thorough

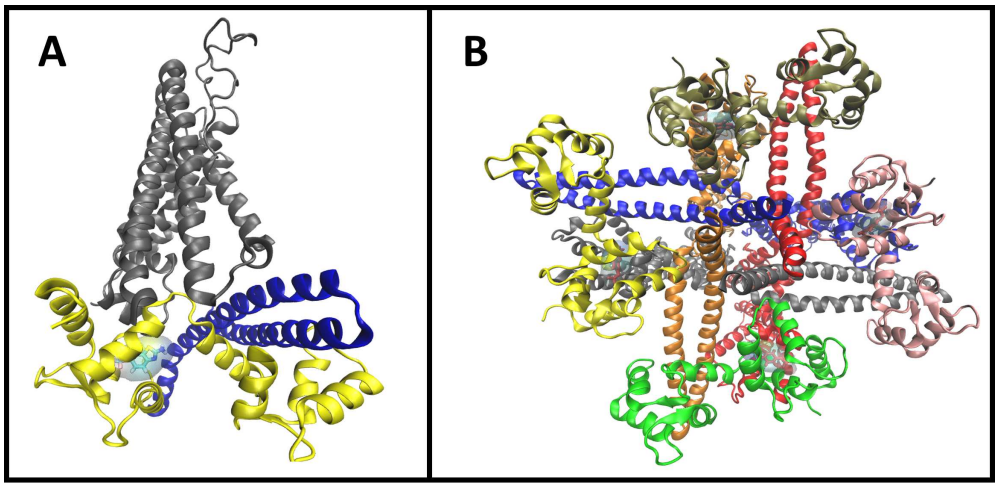


Figure 5.2: Graphical representation of the homology model built for the binding pocket of RIL in the channel SK4. **A**, insight on the binding pocket of RIL (cyan surface), between the N-lobe of Calmodulin (yellow) and the voltage sensor domain of the channel (grey). **B**, bottom view of the homology model of the SK4 channel in the presence of RIL, built from the experimental binding pocket of the SK2 channel, for which the full channel structure is not available.

description of the technique, which is given as an appendix.

5.2 Simulation details and preparation

All-atom molecular dynamics simulations of the full SK4 and $K_V7.2$ channels were performed both in the presence and in the absence of the drugs, using the apo ones as control systems (Figs. 5.3 and 5.4). Each simulation will be run in triplicate, so we will obtain a total of 15 trajectories.

Starting coordinates of $K_V7.2$ were taken from its APO full channel structure (PDB ID: 7CR3 [34]), the one in the presence of ZTZ (PDB ID: 7CR4 [34]) and the one in the presence of RTG (PDB ID: 7CR7 [34]). Since these structures does not contain calcium ions, each N-lobe of calmodulins was loaded with two ions. Similarly, starting structure of SK4 was its full channel structure (PDB ID: 6CNN [76]), in this case N-lobe of calmodulin is fully loaded with two calcium ions and the C-lobe was partially loaded with one. Missing residues were modeled with Swiss model (https://swissmodel.expasy.org/), an online server that reads a sequence of amino

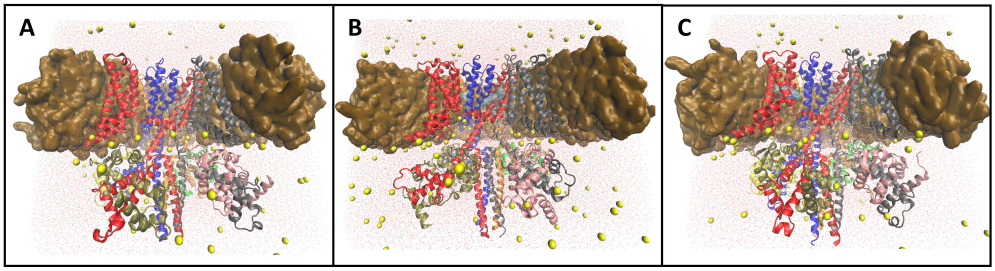


Figure 5.3: Graphical representation of the full channel simulations performed, lipds are represented as a brown surface; water solvent molecules, as red dots; dissolved ions, as yellow spheres, and the channel-drug complexes and drugs hold the format of Fig. (5.1), for **A**, apo run of $K_V7.2$ **B**, $K_V7.2$ in the presence of RTG and **C**, $K_V7.2$ in the presence of ZTZ, they will be referred to as $K_V7.2$_APO, $K_V7.2$_RTG and $K_V7.2$_ZTZ.

acids and aligns it using as templates all proteins that have similar sequences [77, 78]. In our case, only a couple of loops were missing and therefore the templates used by the program were the full channel structures, so our structures were not altered in this step.

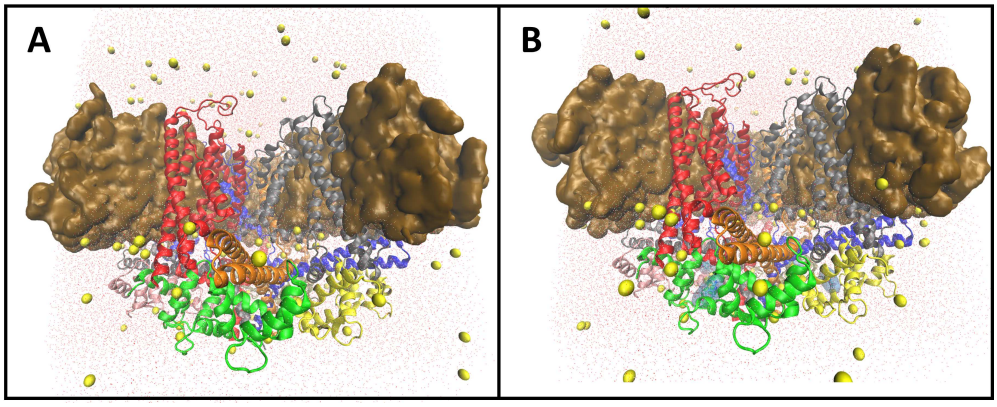


Figure 5.4: Graphical representation of the full SK4 channel simulations performed. Lipds are represented as a brown surface; water solvent molecules, as red dots; dissolved ions, as yellow spheres, and the chanel-drug complexes and drugs hold the format of Fig. (5.1), **A**, apo run of SK4 and **B**, in the presence of RIL. They will be referred to as SK4_APO and SK4_RIL, respectively.

To continue with the systems preparation, the CHARMM-GUI online server (https://www.charmm-gui.org/) was used [97]. All channels were inserted into a 130 Å-long bilayer of POPC lipids except for the $K_V7.2$ channel in presence of RTG which

was inserted into a 150 Å-long bilayer [101], since it is in a different conformation that makes it wider in space. The systems were then solvated using TIP3P water [62], selecting the option of pore water generation of step 1 of CHARMM-GUI. Also a 0.1 M KCl concentration was added in this step. Gromacs inputs for the equilibration and production runs were also generated with CHARMM-GUI [102].

Finally, to avoid potential instabilities due to the random distribution of water molecules by CHARMM-GUI, we removed the ones that lay within 2 Å of the protein or within 8 Å of the membrane lipids.

The chosen parameters were based on the CHARMM force fields, the CHARMM-36 force field was used for protein [21] and lipid atoms [103]. For the drugs, however, there were no parameters available and we produced them using the CHARMM-GUI ligand reader and modeller[1] [104].

As stated before, there is no structural information on the RIL binding pocket in SK4. Therefore, we performed a docking calculation to construct the initial structure, taking the binding pocket of SK2 (PDB ID: 5V02) [3] as a starting point, into the full channel structure of SK4 (PDB ID: 6CNN) [76], by means of the Schrodinger software [105]. For such calculations, RIL was prepared with LigPrep [106] and the ion channel with the protein preparation wizard [107], in this last step, potassium and calcium atoms present in the structure were removed, as well as POV and LMT ligands, which were located where the protein interacts with the membrane, and thus are of no interest for this docking simulation, as they are not part of the binding site we are interested in exploring. Finally, the docking calculations were carried out with Glide [108]. Details of all the docking steps can be found in the supplementary information github page.

To equilibrate the systems, the standard procedure of CHARMM-GUI was followed. It consists of 6 steps, all of them in the NPT ensemble with constraints applied to the membrane and protein. These constraints are gradually decreased in each step. Following reference [100], we extended the last one to 3 ns; since the systems are quite large, they might need more time to stabilize its root mean square deviation and, most importantly, we wanted to be positive that the systems were perfectly equilibrated, as the simulations are computationally very costly. Afterwards, three independent replicas of 500 ns were produced for each system. All simulations were performed with GROMACS 2016.4 [16].

[1] All force field parameters can be found in the "toppar" folders inside each simulation's folder of the supplementary information repository.

For this set of simulations, supporting information is provided in its supplementary information github page [109]. For each run, input files for all simulation steps are provided. In the case of the last thermalization step, the RMSD of all protein atoms is also provided, as a proof that extending this step resulted in properly thermalized systems[2]. In the case of the production runs, together with the final snapshot of each replica a plot with the RMSF of each amino acid throughout the simulation can be found. In those plots, the different domains have been projected on the background to ease the visualization[3]. Also RMSD of the mambrane domain, CaM and C-terminal domain (CTD) can be found[4]. It is interesting that, in general, the CTDs of the $K_V7.2$ channels have more mobility than the SK4 ones, mainly because of the swinging movement of helix hC. In the RMSD and the RMSF plots of the production runs, the error plotted is the standard deviation from the four monomers of the channels. It has to be underlined that, due to the symmetry of the channels, magnitudes will be computed for each monomer. As each system is simulated in triplicate, we will essentially have 12 monomers simulated for 500 ns each.

5.3 Riluzole's pocket characterization

A detailed characterization of the binding pocket of RIL will now be presented. Among the three drugs studied, it is the only one whose full channel structure is not experimentally resolved. Therefore, a full characterization of the pocket obtained from the simulations is certainly valueble.

The first question addressed was whether RIL is actually stable in this pocket. To do so, its root mean square fluctuation (RMSF) was computed, yielding a value of (0.19 ± 0.03) nm. This means that, in average, the standard deviation of the positions of all atoms belonging to the drug is 0.19 nm[5], which is a typical value due to thermal fluctuations. The error of the magnitude comes from the standard deviation over all four RILs that a channel can admit and over the three runs performed. After the stability of the drug in the pocket was settled, we can define the elements of the pocket,

[2]Follow the links for RMSD during last equilibration step for $K_V7.2$_APO $K_V7.2$_RTG $K_V7.2$_ZTZ SK4_APO SK4_RIL

[3]Follow the links for RMSF during production for $K_V7.2$_APO(1,2,3) $K_V7.2$_RTG(1, 2,3) $K_V7.2$_ZTZ(1,2,3) SK4_APO(1, 2,3) SK4_RIL(1,2,3)

[4]Follow the links for RMSD during production for $K_V7.2$_APO(1,2,3) $K_V7.2$_RTG(1,2,3) $K_V7.2$_ZTZ(1,2,3) SK4_APO(1,2,3) SK4_RIL(1,2,3)

[5]Since the channel can diffuse through the membrane, center of mass translations were subtracted from the RMSF calculations.

namely the linker S4S5a and the N-lobe of CaM (CaM N-lobe). A snapshot of the pocket is showed in Fig. (5.5, A) with each α-helix labeled, and each residue is colored according to its probability contact with RIL. The numerical values of these probabilities together with their standard deviation over monomers and runs are provided in (Table 5.1). Some of the CaM residues interacting with RIL have been stated to suffer side chain rearrangements when bound to the SK2 CaMBD [3], namely (Leu-32, Met-51, Glu-54, and Met-71).

Linker S4S5a		
ResID	P	ΔP
Ser-181	0.95	0.03
Ile-182	0.15	0.05
Ala-184	0.994	0.004
Leu-185	0.9998	0.0003
Gln-187	0.7	0.2

CaM N-lobe		
ResID	P	ΔP
Phe-19	0.99	0.02
Ile-27	0.9	0.2
Leu-32	0.9	0.2
Met-51	0.9	0.3
Ile-52	0.7	0.3
Glu-54	0.6	0.2
Val-55	0.99	0.02
Ile-63	0.97	0.04
Phe-68	0.99	0.02
Met-71	0.999	0.008
Met-72	0.5	0.2
Lys-75	0.92	0.01

Table 5.1: List of residues that are in contact with riluzole in more than 10% of the frames, the chosen contact length is 4.5 Å. P is the mean value of the contact probability and ΔP is the standard deviation over monomers and runs.

In order to investigate if the presence of the drug alters the behavior of the residues in the pocket, we computed the root mean square fluctuation (RMSF) of each residue of the CaM N-lobe, which is plotted in Fig. (5.5, B) and of each residue of the linker S4S5a, showed in Fig. (5.5, B - inset), in both cases in the presence and in the absence of the drug. As usual, the error is the standard deviation over monomers and runs. For most CaM N-lobe and S4S5a linker amino acids, the computed error bars of the RMSF of the holo and apo simulations overlap, indicating that the presence of the drug does not cause any substantial differences. However, for residues 54 to 64 of the CaM N-lobe, we see a statistically significant reduction of their mobility. It turns out that this segment, depicted in Fig. (5.5, C - yellow), is very relevant, as it is one of the loops that can accommodate a calcium ion in the starting structure and throughout the simulation. To have an alternative description to validate this difference in

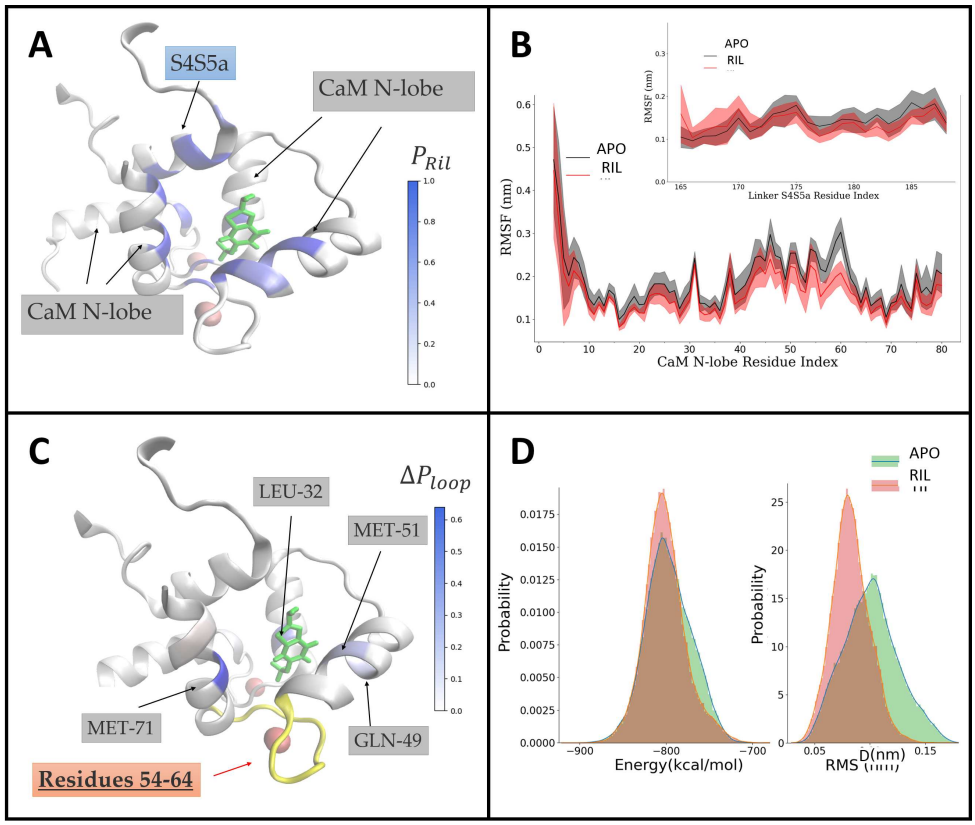


Figure 5.5: Characterization of the binding pocket of riluzole in the SK4 channel.**A**, Contact probability of RIL (green) within its binding pocket (P_{Ril}). Helices belonging to the CaM N-lobe or the linker S4S5a are labeled. **B**, RMSF of all amino acids belonging to the binding pocket depicted in **A**, where the S4S5a linker is showed as an inset. Values of apo runs are connected by black lines and RIL values, by red lines. The standard deviation over the four monomers of each of the three runs is showed as a dashed fill. **C**, difference of interaction probability of residues 54-64 (yellow) with the rest of the binding pocket, $\Delta P_{loop} = P_{loop}^{APO} - P_{loop}^{RIL}$, so that the blue color means that the residue interacts more with the yellow loop in the apo runs. **D**, Histograms of the interaction energy of Ca^{2+} with nearby protein residues or RIL if present (left) and the RMSD with respect to the minimun energy frame of apo and Ril runs (right), information of all monomers of all runs were used to construct the histogram.

mobility, we computed the contact probability of each amino acid of the pocket with the loop, P_{loop}. The difference between the probabilities of the apo runs and in the presence of the drug $\Delta P_{loop} = P_{loop}^{APO} - P_{loop}^{RIL}$ is showed in Fig. (5.5, C). There were no residues with a higher contact probability with the loop in the presence of RIL and residues Leu-32, Gln-49, Met-51 and Met-71 yielded values of $\Delta P(Leu\text{-}32) = 0.4 \pm 0.3$,

$\Delta P(Gln\text{-}49) = 0.2 \pm 0.3$, $\Delta P(Met\text{-}51) = 0.2 \pm 0.3$ and $\Delta P(Met\text{-}71) = 0.6 \pm 0.2$, meaning that, in the presence of the drug, the loop interacts less with these residues, which can be interpreted as the loop being less flexible to make the contacts that it usually does in the absence of the drug.

We wanted to correlate differences in flexibility with other quantities that would allow us to state a hypothesis of the effect of the drug in the channel. With this objective, we computed the interacting energy of the Ca^{2+} bound to the loop under study with all protein residues and RIL (in the runs that is present) for all time steps of all simulations, by means of the NAMD energy plugin [15] implemented in VMD [33] with the CHARMM36 force field [21]. A histogram of the computed energies is showed in Fig. (5.5, D - left). It is interesting that, for the runs in the presence of RIL, the histogram shows a Gaussian-like distribution with a minor shoulder around high energies (~ -750 kcal/mol). This Gaussian shape is expected as the interaction energies are dominated by the thermal fluctuations of the system. In contrast, for the apo runs, a bigger shoulder is found around ~ -750 kcal/mol. This increment of probability for higher energy terms could be revealing and would imply that the ion would be more stable in the presence of RIL. Finally, we computed the RMSD of the atoms that are less than 5 Å away from the ion and depicted it as a histogram in Fig. (5.5, D - right), using the frame of smallest energy from the three runs of each system to set the reference for the RMSD calculation. As before, there is a marked shoulder around high deviations for apo. To connect the two ideas that we have introduced, this is, that RIL decreases the mobility of the loop and increases energetical stability, we compute the Pearson correlation coefficient of the energy and the RMSD, defined as:

$$corr(E, RMSD) = \sum_{j}^{monomers} \sum_{i}^{frames} \frac{(E_{ij} - \overline{E})(RMSD_{ij} - \overline{RMSD})}{\Delta E \Delta RMSD} \tag{5.1}$$

yielding a result of 0.52 for APO and 0.32 for Ril, supporting the idea that there is a connection between calcium stability and mobility of the loop. The smaller correlation for RIL can be another symptom that the thermal effect is more relevant in the presence of the drug. These results also suggest that the molecular mechanism of the drug would be to stabilize the loop in order to enhance the stability of the bound calcium ion, which is consistent with RIL acting as a channel activator by enhancing its affinity for calcium ions [3]. This hypothesis agrees with the fact that amino acids Gly-25, Gly-

61, Gly-98, and Gly-134, each belonging to one CaM loop holding a calcium ion, were experimentally found to be perturbed in the [CaM-SK2 CaMBD-RIL] complex (PDB ID: 5V02) [3], leading the authors to think that these glycines act as RIL-mediated hinges that make subtle changes increasing Ca^{2+} affinity.

5.4 Network Analysis

5.4.1 Introduction

With the aim of studying allosteric regulations within proteins (i. e. long-range interactions), graph theory has been drawing an increasing attention over the last two decades [110–113]. Graph theory is a mathematical theory that consists in extracting information out of a graph, which is a set of nodes connected by edges. These nodes will usually have structural features of the system topology; for example, they can be the α-carbon of one amino acid. Notice that this definition of the node implies a simplification of the system as all information corresponding to one amino acid is contained in one node, as in the case of coarse grained force fields. Finally, edges will take into account the connection between the nodes of our system. These connections are translated into a square matrix, usually called the adjacency matrix A_{ij}, whose matrix elements are equal to zero if nodes i and j are not connected, and a positive value otherwise, so that A_{ij} accounts for the strength of connections between nodes but also if they are connected at all.

The main reason for introducing graph theory to the analysis of our simulations is the concept of node centrality. Essentially, it assesses the importance of the nodes within the network. Several ways of defining centrality have been proposed [114] but, when applied to allosteric regulation, the common method is to make use of betweenness centrality, which has been found to distinguish relevant residues in allosteric signaling [115–117]. Betweenness centrality, also known as geodesic centrality, is the fraction of all possible shortest paths connecting any pair of nodes of our system that cross a given node. This magnitude provides information about which nodes are more important for general propagation of signals, a node that has low betweenness centrality will be in the outskirts of the topology and will be less involved in transmitting any signal. Therefore, centrality measures are capable of translating all the information contained in the edges of the graph into a single number for each node, drastically reducing the amount of data that will be analyzed.

There are some systems in which the start and end of the signal is known and the interest lies in how this signal is transmitted. This is the case of ion channels, in which the pore is allostericaly connected to the C-lobe of calmodulin, in the case of Ca^{2+}-gated channels where CaM acts as a Ca^{2+} sensor, or to the voltage sensor domain, in the case of voltage-gated channels. In these cases, the knowledge of the starting and ending nodes of the signal changes the recommended definition of centrality, we will be interested not only on the shortest paths but also in all nodes that might have a contribution for propagating the signal. Otherwise we might have instabilities when applying the method to molecular dynamics as different replicas might yield different shortest paths and there might also be fluctuations over time.

Recently, current-flow betweenness centrality (CFB), a new methodology for computing betweenness centralities, has been introduced [117]. In this framework the start and end of the signal need to be specified and its foundations are based on the analysis of electrical circuits [118]. Additionally, it has been proved that this measure is the asymptotic limit of a random-walk-based method [119]. The advantages of this method with respect to the geodesic shortest path is that it has been proven to be more robust, computationally less expensive and it is a cut-off free method [117], being the latter an improvement with respect to suboptimal pathway calculations [120].

Given that graph theory only requires a set of nodes and the connections between them, the abstraction of the protein into a graph can be very varied, as reviewed in [112]; for proteins, most works consider that each node is an amino acid, but for the weights of the connections there is no consensus in the definition. Some works make use of unweighted graphs (all edges are equal to 1), and, in others, the weights are the number of atoms that are in contact between the two nodes, the inverse of the averaged distance, or the average contact energy, among others. Finally, Network Analysis will be the application of graph theory to compute the current-flow betweenness centrality of the nodes of our system. In the following paragraphs, we will see the exact definition of our graph and the mathematical foundations behind the calculations.

5.4.2 Definition of the network

For the definition of our network, we followed the methodology presented in [100, 121]. Each amino acid refers to one node and lipids will be separated into three (i. e. one node for the hydrophilic head and two for the hydrophobic tails). The

adjacency matrix will be the product of the contact map and the mutual information of the fluctuations of the positions of the pair of nodes, so that both structural and dynamical information of our simulated system is introduced into the network.

The contact map will contain structural information of our system and will help us to discard each pair of non-interacting nodes, making the calculation of the graph easier. To determine it, we will not use a binary value; instead, for distances smaller than the cutoff $c = 4.5$ Å, the probability will be set to one but for bigger distances a smoother probability distribution $K(d)$ will be used so that the probability goes smoothly to zero, as it has been proven to be more reliable [117]:

$$K(d_{ij}) = \frac{exp(-\frac{d_{ij}^2}{2\sigma^2})}{exp(-\frac{c^2}{2\sigma^2})} \tag{5.2}$$

where d_{ij} is the distance between atoms i and j. This distribution is used up to a cutoff of 8 Å, the standard deviation of the gaussian σ is 1.38 Å. All the dynamical information of the simulation is translated into the contact map as the value of $K(d_{ij})$ is averaged over all time steps. Since we are performing this average, the appropriate time scale for these technique will be around the hundreds of nanoseconds, as no major conformational changes are sought but statistical robustness is required. Otherwise, we might obtain unrealistic contact maps.

The contact map determines the connections in our network, but we also want to add the information about the correlation between the fluctuations of each pair of nodes. This is achieved by computing the mutual information matrix of the atomic fluctuations. The mutual information of two random variables contains the information can be obtained from one variable by knowing the other one. In other words, how correlated the two variables are. In our case, we will compute the matrix of mutual information M_{ij}, for two nodes (or amino acids) i and j:

$$M_{ij}(x, y) = \int \rho_{ij}(x, y) ln \frac{\rho_{ij}(x, y)}{\rho_i(x)\rho_j(y)} dxdy \tag{5.3}$$

where $\rho_i(x)$ and $\rho_j(y)$ are the probability density distributions of nodes i and j fluctuating by x and y around their resting positions, so using x and y is a compact notation for displacements $\mathbf{x} - \overline{\mathbf{x}}$ and $\mathbf{y} - \overline{\mathbf{y}}$; similarly $\rho_{ij}(x, y)$ is the joint probability density, this is the probability that node i has a fluctuation x and node j has a fluctuation j. The mutual information is defined as a non-negative value. Notice that, in particular, if variables x and y are independent, the joint probability distribution is equal to the

product of the marginal distributions, $\rho_{ij}(x.y) = \rho_i(x)\rho_j(y)$, so that the mutual information is zero when the variables are not correlated. The mutual information can be written in terms of the statistical entropies (H), of the variables :

$$M_{ij}(x, y) = H_i(x) + H_j(y) - H_{ij}(x, y) \tag{5.4}$$

Where the entropies are defined as $H_i(x) = -\int \rho_i ln \rho_i(x) dx$, this relation is important since, in practice, the mutual information will be computed in terms of the entropies, as it is computationally less costly. Its mathematical derivation is attached as part of Appendix A, and further information on how Eq. (5.4) is computed in practice can be found in [121]. Finally, the adjacency matrix will be defined as the product of the averaged contact map and the mutual information: $K(d_{ij})M_{ij}$ for nodes $i \neq j$ and 0 for the diagonal elements.

5.4.3 Current flow betweeness centrality

To characterize the allosteric pathways in our system, we will compute the Current Flow Betweeness centrality (CFB). It will help us illustrate which amino acids are more important to transmit a given signal. The main idea is that we are computing all possible paths connecting sources and sink, characterizing how much information flows through each node and pouring it into the centrality measure. In Appendix A a thorough derivation of its mathematical expression [118] can be found, here we just write its expression:

$$\tau_i = \frac{1}{2|S_0|} \sum_{s,j} |p_i(s) - p_j(s)| A_{ij} \tag{5.5}$$

In sort, we compute the information current passing through node i, coming from all its adjacent nodes j, and this is repeated for all sources s, so the final value is normalized by the total number of sources S_0. Note that the information flow is treated as an electrical current, so the previous equation is like Ohm's law, in which the electrical current is the potential difference between nodes times the conduntance (this is, the inverse of the resistance).

As introduced before, in the particular case of ion channels, the sources will be the gating charges of the voltage sensor domain in the case of $K_V7.2$, or the C-lobe of calmodulin for both studied channels, since the SK4 channel is voltage insensitive [30]. The sink will always be the pore, following reference [100]. The identification of these residues for the $K_V7.2$ and the SK4 channels was carried out following references

[34] and [76], respectively. these residues can be found at Table 5.2. In the following section, we will start by reviewing how the results are provided. A more detailed explanation of the calculations and the code used can be found in [122].

Table 5.2: Residues used to define the sources and sinks of the studied signals.

Sources	**K_V7.2**	**SK4**
VSD gating charges	198, 201, 207, 210 (Arg)	-
CaM C lobe	80-148	80-148
Sinks	**K_V7.2**	**SK4**
Pore	314 (Ser)	282 (Val)

5.5 Current flow betweenness centrality calculations

In this section, the calculations of the CFB for all simulations are presented , for both the K_V7.2 and the SK4 channels. Before we start, let's comment on how this magnitude is represented in the following figures. Since the channels have four identical monomers, the signal will be defined to begin at the source residues of the four monomers and to end at the four pore residues. From the definition of the supply vector (see Appendix A), each term of the sum in Eq. (5.5) comprises one source residue and all sinks, so when we are computing the signal starting in the CaM C-lobe, for example, we are averaging over the signals starting in all its residues of each monomer and all arriving at the four pore residues.

The calculation of a signal will then give the CFB of all amino acids of the tetramer, but, for simplicity, in the following plots these are averaged over the monomers so that only one monomer is presented.

It also stands out that the pore residues have a zero CFB value in all plots, as stated in the mathematical derivation of Appendix A, for a CFB calculation both the pore and the sources are given a value of zero, and the pore is the same in all terms of Eq. (5.5) so its value is always zero when averaging between sources and monomers. Conversely, each term of Eq. (5.5) refers to a signal starting in that source and we set that value to zero, but different terms referring to nearby sources will yield a CFB different from zero, so that, after averaging, the total CFB is different from zero.

Finally, we will also compute the averages over the three different replicas that we have for each system. Therefore, the depicted errors as will be the standard deviation over the different runs and monomers.

As we can see in figure Fig. (5.6, A), we represented the current flow betweenness centrality of all amino acids of the channel and calmodulin, ordered by sequence. For clarity, we have joined the CFBs of the different residues with solid lines, and also we have illustrated the different domains by coloring the background, adopting the following color code: transmembrane helices S1-S6 are in blue, intracellular helices hA-hB-hC are in yellow and calmodulin lobes are in green. Before we comment on the results, let's address the general features of these graphs for the sake of clarity, taking advantage of the fact that the mathematical derivation of the magnitude is presented in Appendix A.

5.5.1 General features

CFB can be seen as the propagation of an information current, modeled using circuit theory. Therefore, we used current conservation laws in the mathematical derivation, meaning that the current that flows into one node is equal to the one flowing out of it, so that nodes located near an important node will benefit from it and will necessarily have greater centralities. In other words, the magnitude of the centrality also accounts for the location of the node. One example of this is that, in all plots, residues near the sources or sinks have in general high centrality. This is because the information flows from sources to sinks and is conservative, so it will spread from the sources and will gather around the sinks, increasing the centrality of the nodes nearby. This introduces an overestimation near the sources and sinks and, therefore, it has to be taken into account. Unfortunately, it is very difficult to estimate a cut-off from which this effect is negligible.

A common feature of the graphs is that, in general, the CFB is higher in amino acids belonging to helices. This can be due to the fact that secondary structure makes the motion of amino acids to be more correlated, making them excellent transmitters of information flow. Conversely, unstructured linkers have more freedom to fluctuate, resulting (in general) in a lower mutual information with the surrounding residues, blocking the information flow signal.

Another interesting aspect is that, in helices, the graphs tend to have a saw-tooth-like shape, if we look closely, it can be noticed that it has the same periodicity of the helical turns, as in this case the information follows two paths: it travels following the sequence but also straightly jumping through the hydrogen bonds, the latter effect is the one responsible of the saw-tooth like shape.

In general, we will compare the obtained CFBs with the ones of the control systems, since it will be difficult to rationalize a given CFB value and also the plots are not easy to read.

After this introduction to the CFB calculations using the Network Analysis methodology, we are in position to present the obtained results.

5.5.2 Apo $K_V7.2$

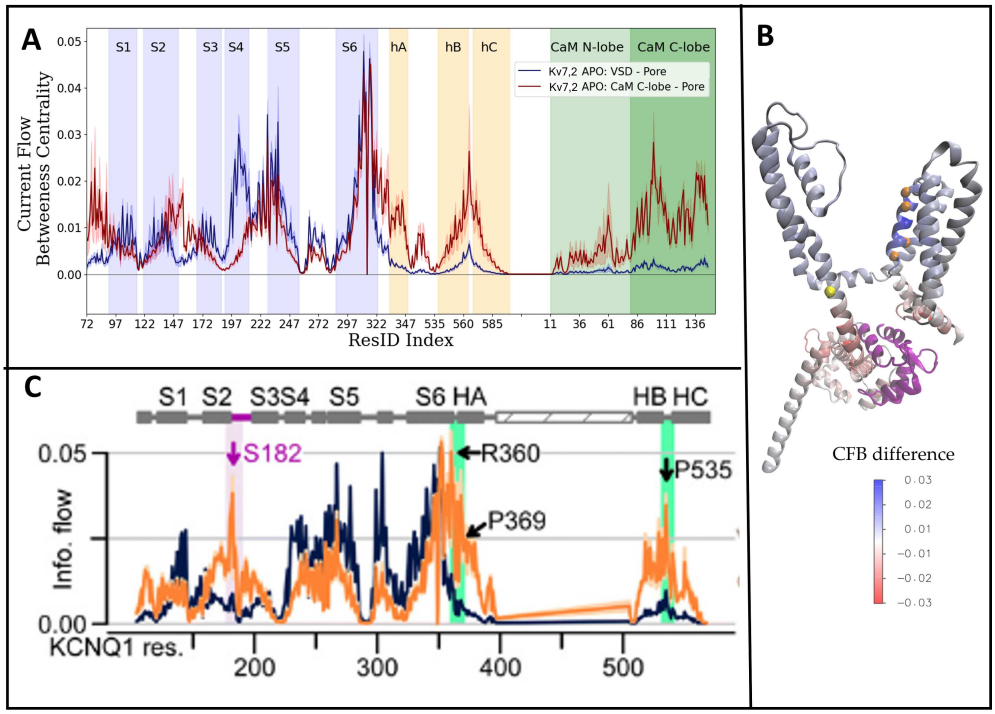


Figure 5.6: Results obtained from the Network Analysis for the $K_V7.2$ apo runs. **A**, mean CFB obtained for each residue for all apo runs, for the signal traveling from the VSD gating charges to the pore (blue) and from the CaM-C lobe to the pore (red). The standard deviation over different runs and monomers is showed as dashed fills. The domains of the channel are colored in the background. **B**, Difference between the CFB of the VSD-pore signal and the CaM C-lobe-pore signal. The gating charges are depicted as orange balls; the pore, as a yellow one and the CaM C-lobe is colored in purple. **C**, data extracted from [100] in which the authors performed the same calculations as in **A** but for the $K_V7.1$ channel. Some centrality hubs are emphasized and, at the top, the domain to which each amino acid belongs is shown.

Let us analyze now the CFB results, considering the general comments previously

made. To begin with, for the $K_V7.2$ channel in the absence of drugs, the signals starting at the gating charges of the VSD and at the CaM C-lobe have been represented in Fig. (5.6, A: blue) and in the Fig. (5.6, B: red), respectively. Looking at the first one, we can have a feeling of how the signal is transmitted from the VSD gating charges to the pore. Firstly, there is a great hub of centrality around helices S4 and S6, as the sources are located in S4 and the sink in S6. Also, S5 exhibits a high centrality, and to a lesser extent S1, S2 and S3. It is very important to know the three-dimensional arrangement of these helices and the positions of the gating charges to conclude that, roughly speaking, the flow follows mainly three paths: through S5', where the apostrophe stands for the fact that it belongs to the contiguous monomer (arginine 198 and 201 are in contact with S5'), through the S4S5 linker (arginine 207 and 210 are closer to the S4S5 linker) and, to a lesser extent, through S1 (arginine 201 faces also the S1 helix).

In the case of the signal flowing from the CaM C-lobe to the pore, showed in Fig. (5.6, B: red), it can be noticed that the centralities differ considerably with respect to the VSD-pore signal; as expected, there are centrality hubs in the CaM C-lobe (source) and near the pore (sinks). However, this time there is an increase in the centrality of the residues near S6 and hA helix, as part of the signal is now traveling through the the cytosolic domain, while the rest goes through the transmembrane domain as in the previous case. Looking at the cytosolic domain, there are two big hubs of centrality. One is hA helix together with its linker to the pore domain, and the other is the linker between hB and hC, which also lies near the pore.

To elucidate the disparity of the two signals, we can compute their difference for each amino acid; results for the CFB of the VSD-pore signal minus the CaM C-lobe-pore one are showed in Fig. (5.6, B). In this case, the results have been projected onto the three-dimensional structure of a monomer. It is clear that the whole cytosolic domain is influenced mostly by calmodulin and the transmembrane domain by the VSD charges, there are some exceptions in the latter though, like the residues before S1 and the region near the S2S3 linker, which lie near CaM.

Our results are similar of those published in [100] for the Kv7.1 channel, showed in Fig. (5.6, C), whose structure is very similar to that of Kv7.2 [13]. In fact, all the S1-S6 and hA-hC domains are conserved. Note that, in Fig. (5.6, C), the centralities of CaM residues are not showed. When comparing our centralities with theirs, it stands out that their pore helix, a small helix that lies between S5 and S6 in the

sequence, has a bigger centrality, both when the signal starts in the VSD or in the CaM C-lobe, which could be an indicator that the pore helix is more important for the $K_V7.1$ channel. Furthermore, in the case of the of the signal starting from the CaM C-lobe, they observed a high peak in the S2-S3 linker, which lies very close to the CaM C-lobe. Its absence in the $K_V7.2$ channel and the fact that our centralities at the residues before the S1 helix (also close to the CaM C-lobe) are higher, might indicate that the communication between CaM and the transmembrane domain could be different in the channels. However, this does not seem easy to check experimentally, and the cause could just be that the experimental structures of the two channels are in slightly different conformations.

5.5.3 Drugs

Continuing with the results obtained, we can comment on how the different drugs affect the current flow betweenness centrality of the channels. As introduced before, one of the key aspects of the selected drugs is the variety of the binding pocket locations (see Figs. (5.1-5.2)), so that, presumably, the effect of the drugs should be different. Let us analyze the effect of the different drugs separately.

ZTZ240 (ZTZ)

The ZTZ binding pocket was showed in Fig. (5.1, C and D), it is situated between the four helices that constitute the VSD. When the signal starts in the VSD gating charges, it has a CFB of 0.016 ± 0.004, which is similar to that of residues belonging to the S1, S2 or S3 helices. All centralities for this signal are depicted in Fig. (5.7, A). In general, when comparing with the apo curve, we can see that the plots are very similar, but there are many small differences in the height of some peaks. The one that we found that could be the most important is that at Phe-248; the gain of betweenness of this residue in the presence of the drug could indicate that the drug enhances the communication with the pore through this amino acid. To elucidate this, we computed the contact probability of Phe-248 with the nearby residues, observing a noticeable increment with several residues in the VSD in the presence of the drug when compared to apo channel, they are depicted in Fig. (5.7, B). It stands out that there is an increment with Arg-198 and Arg-201, which are two of the four gating charges, but also with three more residues in the S4 helix, namely Phe-202 and Ile-205. This results suggest that the presence of ZTZ implies a small rearrangement of the VSD so that the communication with S5' is enhanced through the Phe-248 residue, which

would serve as a bridge connecting the two domains.

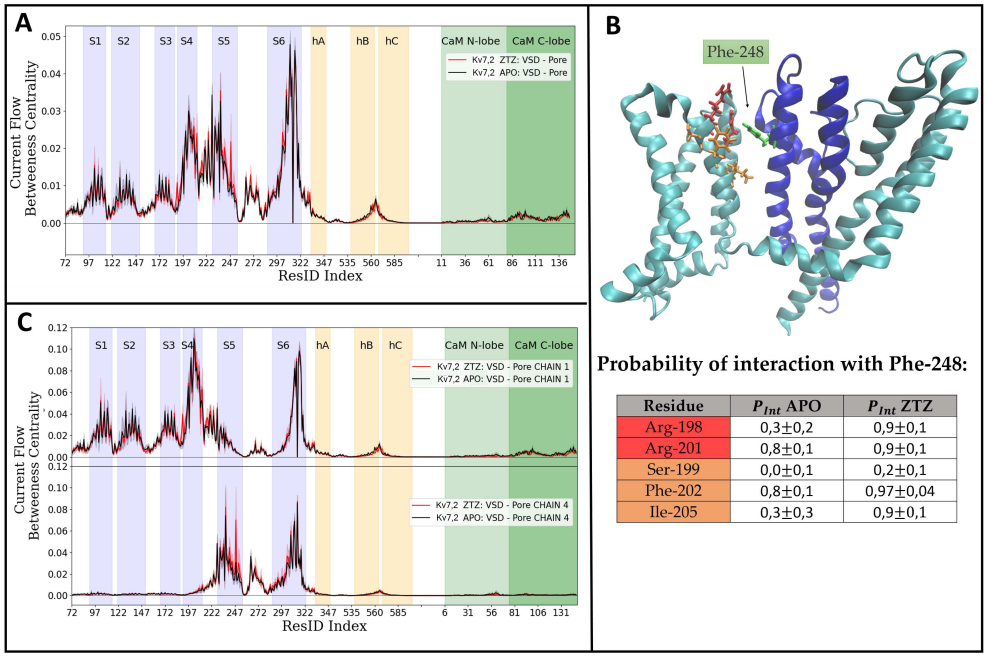
The importance of the communication between S4 and S5' can be illustrated if, instead of including the sources of the four monomers in the sum of Eq. (5.5), we consider them separately, and assign the name of chain 1 to the monomer in which the signal is originated and chain 4 to the adjacent one; chains 2 and 3 will not be represented as their amino acids will almost not exhibit CFB, except for amino acids near the pore, as they lie far from chain 1. Therefore, we can have, as before, 12 signals to obtain the standard deviation but still have the information of how the signals jump between monomers. As can be seen in Fig. (5.7, C), much information that came from the VSD of chain 1, (Fig. (5.7, B: cyan)) jumps to chain 4 (Fig. (5.7, B: dark blue)). This is expected, since this channel exhibits a domain-swapped conformation, which means that its monomers are interlaced.

Going back to the standard type of plot for the signals, when it flows from the CaM C-lobe to the pore we do not observe any major changes with respect to APO worth commenting, it is showed in Fig. 5.8). In this case, the CFB of the drug is 0.007 ± 0.003, smaller than the previous situation. This indicates that the drug would be more involved in the signal that starts at the VSD, which can be related to the position of its binding pocket.

Retigabine (RTG)

The binding pocket of RTG is depicted in Fig. (5.1, A and B), it is located next to helix S6 and facing the membrane. Looking at the results, RTG shows a CFB of 0.022 ± 0.008 for the signal starting at the VSD and 0.015 ± 0.005 when starting at the CaM C-lobe. The VSD-pore signal is showed in Fig. (5.9, A), we do not observe any difference in the transmembrane domain when compared with the apo runs, and the cytosolic domain has low CFB, indicating that in this case almost all information goes through the membrane. Conversely, there are more differences between apo and RTG when the signal starts in the CaM C-lobe, as it can be seen in Fig. (5.9 B): there is a CFB reduction in helix S1 and the previous residues and an enhancement in the cytosolic domain. This would indicate that there is less communication with CaM and for that reason more information would flow through the cytosolic domain.

To rationalize these results we have to comment on a peculiarity of the structure of the channel in the presence of RTG: it exhibits what is called a CTD-straight conformation, depicted in Fig. (5.9, C: red), where CTD stands for C-terminal domain, in which the S6-hA linker is different with respect to the apo and ZTZ structures in Fig. (5.9 C: blue), in presence of RTG this linker no longer exist as helices S6 and



Residue	P_{Int} APO	P_{Int} ZTZ
Arg-198	0,3±0,2	0,9±0,1
Arg-201	0,8±0,1	0,9±0,1
Ser-199	0,0±0,1	0,2±0,1
Phe-202	0,8±0,1	0,97±0,04
Ile-205	0,3±0,3	0,9±0,1

Figure 5.7: Results obtained from the Network analysis for the $K_V7.2$ simulations in the presence of the ZTZ drug. **A**, CFB results of the VSD-pore signal in the presence of the drug ZTZ (red) compared with those of the apo runs (black). **B**, visualization of Phe-248 amino acid (green) belonging to S5' in chain 4 (adjacent monomer, in blue), and some of the interacting residues of S4 in chain A (cyan), namely Ser-199, Phe-202 and Ile-205 (orange) and the gating charges Arg-198 and Arg-201 (red) with a table comparing the interaction probability of these residues with Phe-248 in S5' in the apo channel and in the presence of the drug. **C**, breakdown of the CFB by chains. For this calculation, the CFB has not been averaged over the monomers, chain 1 and 4 refer to the cyan and blue monomers in **B**, respectively.

hA are united into one straight helix (CTD straight). Conversely, in the apo and the ZTZ structures this linker has no secondary structure and forms an angle (CTD bent). This structural change has other peculiarities, as it causes the separation of calmodulin from the VSD. This would explain then the differences observed in Fig. (5.9, B), as CaM is further away from the VSD, it does not interact with the residues before S1, lowering its centrality, and the residues between hA and S6 exhibit an increase of their centrality due to their different secondary structure.

It is difficult to obtain insights on the effect that this drug has on the channel, aside from the conformational changes observed in the experimental structure. Taking a different perspective, we can state that the technique is able to detect conformational

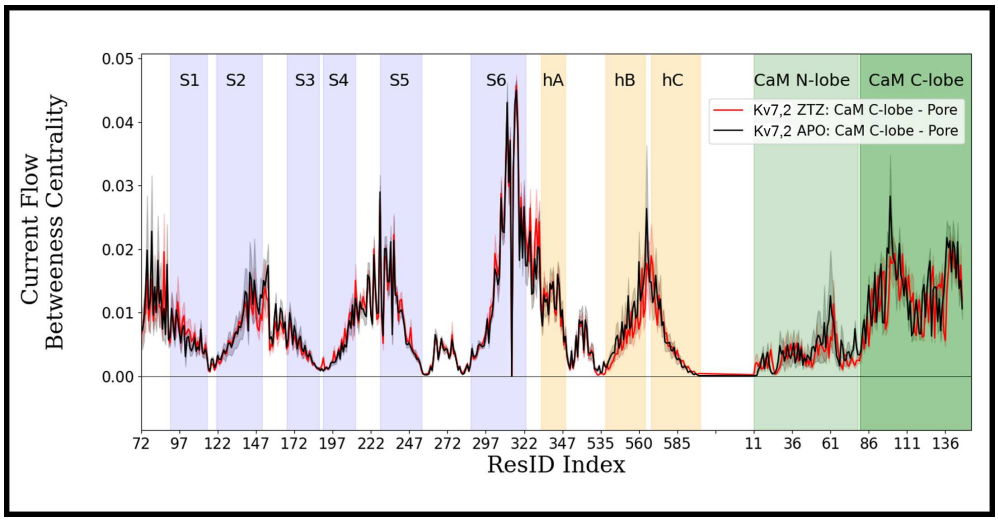


Figure 5.8: CFBs of the $K_V7.2$ and CaM amino acids for the CaM C-lobe-pore signal in the presence of ZTZ (red) compared to the apo runs (black).

changes, even though in this case it was obvious from visualizing the structures.

Riluzole (RIL)

Regarding RIL and the SK4 channel, the pocket modeled according to the experimental information available lies between the CaM N-lobe and the helicoidal linker S4S5a, depicted in Fig. (5.2). As stated before, the SK4 channels have lost the sensitivity for potential differences [30], so only the CaM C-lobe-Pore signal is studied, and is showed in Fig. (5.10, A). The CFB of RIL for this signal is 0.03 ± 0.02, which is quite large considering the centrality of the amino acids belonging to the pocket , and it shows a larger standard deviation than the previous drugs. This made us consider the robustness of the calculations between different runs, which will be commented later on, but before we get to that let us comment on the centralities plotted in Fig. (5.10, A). Several residues in the pocket show a small increment of their CFB. Some of them are in the CaM N-lobe and two are in the S4S5a linker and in the linker connecting this helix to helix S4. However, these increments are not statistically significant as their errors overlap with those of the apo runs. Additionally, there is an increment in CFB on the S2-S3 linker and in the hA helix. The S2-S3 linker can be found in Fig. (5.2, A); it is the gray linker lying very close to the dark blue one (which is the hB helix of the monomer in front). Also, small reductions in all areas that lie in

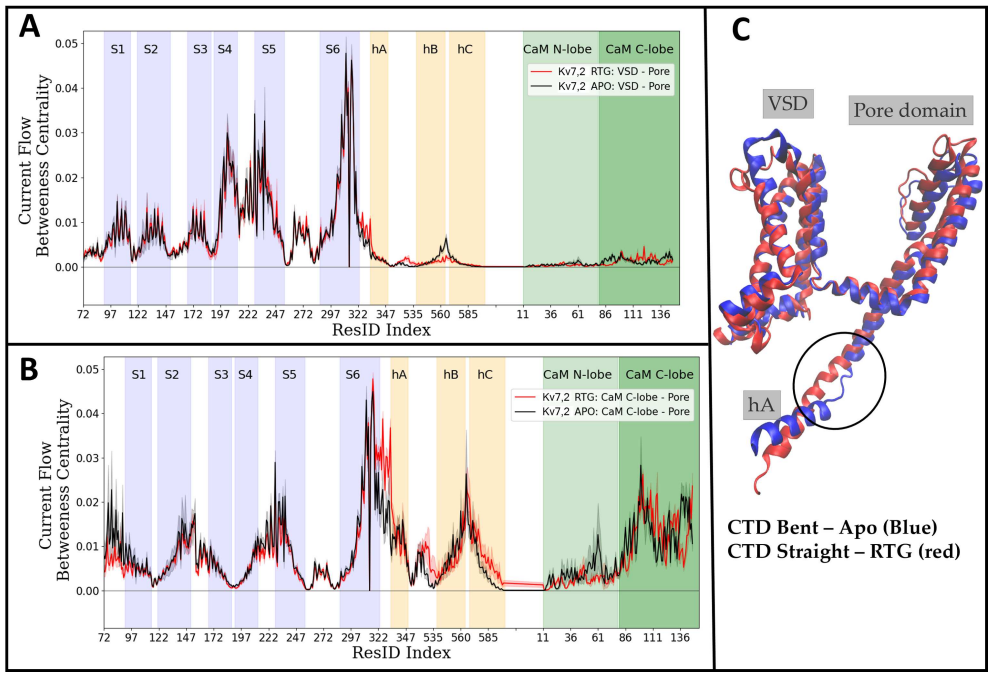


Figure 5.9: Results of the Network Analysis for the $K_V7.2$ channel in the presence and the absence of RTG. **A**, CFB of the $K_V7.2$ and CaM amino acids in the apo (black curve) and RTG (red curve) run for the VSD-pore signal. **B**, CFB of the amino acids of the sequence in the apo run (black curve) and RTG run (red curve) for the signal starting in the CaM C-lobe and ending in the pore. **C**, Three-dimensional starting structure of the apo run (blue), which is in the CTD bent conformation, and the one of RTG (red), which is in the CTD straight one.

the transmembrane domain but closer to the outer part of the cell can be found. For example, the amino acids near the S1-S2, S3-S4, S5-pore helix or pore helix-S6 linkers. These two features could indicate that RIL causes the information to flow following paths involving the cytosolic domain, in particular hA, and minimizing the contributions from amino acids near the outer environment, but this effect is not statistically significant.

In order to have a closer look at the amino acids belonging to the pocket and check the robustness of the calculations, we separated the calculated CFB of each of the three runs in the presence of RIL and compared it to the mean apo value. The S4S5a helix is showed in Fig. (5.10, B: left), there it can be seen that, in general, the Ril CFBs computed at each run are different. For the amino acids belonging to the CaM

N-lobe, showed in Fig. (5.10, B: right), we obtained a similar result, the CFBs values of different runs were generally different, in this case, it is interesting that almost all amino acids that interact with RIL exhibit slightly higher CFB than the mean apo runs, but again this is not statistically significant.

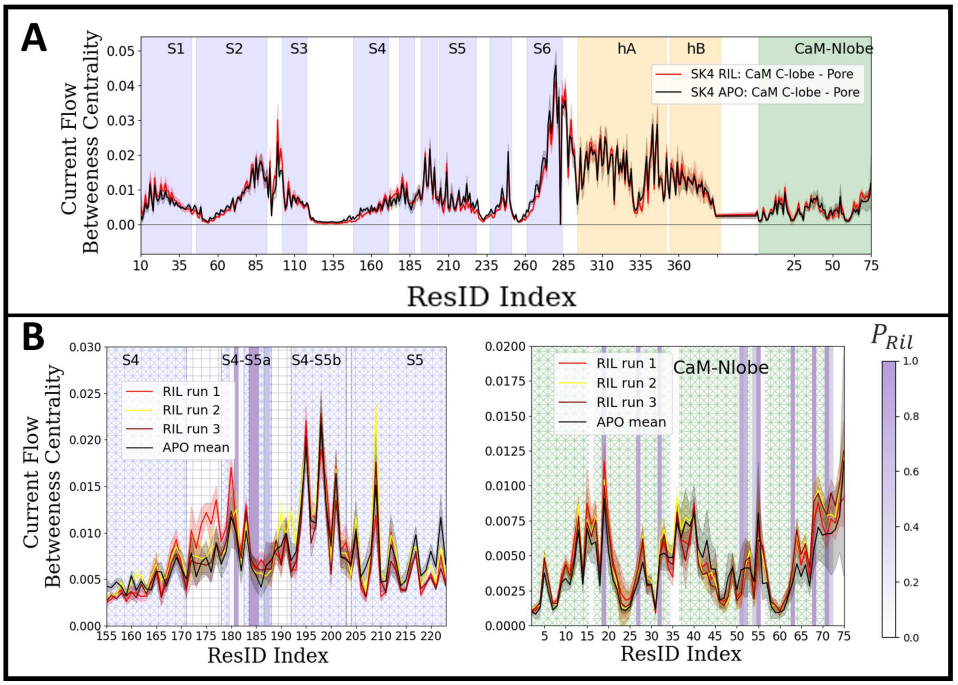


Figure 5.10: Resuls of the network analysis for the SK4 channel in the presence of riluzole. **A**, CFB of the SK4 apo channel and in presence of RIL for the signal flowing from the CaM C-lobe to the pore. **B**, insight of the CFBs in the part of the binding pocket belonging to the channel (left), and to the CaM N-lobe (right); the CFB of each of the three RIL runs are plotted separatedly and compared to the mean apo value. The contact probabilities of the S4S5a residues with RIL is projected in the background with a color scale. A zero probability is represented as pattern-filled surfaces with the same background used in previous figures to label the domains.

5.5.4 Conclusions

In this chapter, molecular dynamics simulations were carried out for the full channel structures of $K_V7.2$ and SK4, in the presence and in the absence of some activators of these channels; the goal was to see if it was possible to elucidate the atomistic mechanisms by wich they enhance the channels' function. Considering the distances of the

binding pockets to the pore, it is clear that the drugs must work allosterically. The huge amount of information obtained from an MD simulation makes it almost impossible to extract meaningful information on these mechanisms by mere visualization, unless a more sophisticated method is used. One candidate methodology for describing the pathways or signals and how these drugs affect them is network analysis, which was reviewed and applied in this chapter. The CFB results obtained can be regarded as the importance of a certain node given a defined pair of hubs that define the sources and sinks of information flow. It has to be underlined that the CFB calculations take into account all possible pathways connecting the sources and sinks and pour all this information into one value per residue, which is a major reduction on the amount of output data to be analyzed.

From the results obtained for the $K_V7.2$ channel in the absence of drugs, the comparison with the analogous channel $K_V7.1$ raised apparent differences between their communication with CaM and in the importance of the pore helix. It is difficult, however, to translate these CFB differences into measurable observables, but one possible explanation is that their experimental structures might be in slightly different conformations.

Further comparisons of our CFB of the apo channels with those obtained in the presence of drugs were carried out. In the case of ZTZ, the small conformational changes in the VSD induced by the drug cause a higher CFB in Phe-248. Also, the contact probabilities obtained suggest that the increment in CFB of this amino acid is because it serves as a bridge connecting the VSD gating charges with S5' (this is, the S5 helix of the adjacent monomer). This could result in the enhancement of voltage sensitivity caused by the drug, however, further experiments should be carried out to confirm this hypothesis.

For the case of RTG, it was not possible to extract any information, since the differences in the CFBs came from the different conformations of the starting structures. Since the binding pocket of RTG involves the membrane, the effect of the drug could have something to do with it, but it was not possible for us to extract any meaningful information from the CFBs of the membrane, even though they are explicitly included in the calculations.

In the case of riluzule, the small differences obtained were not statistically significant. However, as there were small increments in several channel residues, we came up with the idea of computing the average RMSF of the amino acids of the pockets and ended up with the characterization of the RIL pocket and its possible atomistic mechanism, presented at the beginning of this chapter.

To conclude, this technique can be useful to detect small conformational changes or mechanisms from the simulations, but, in our case, the effects of the drugs were not clear. Even though an all-atom model was used, it contains several approximations, when comparing the simulations with and without the drug, we were only changing few tens of atoms (one node) out of three hundred thousand (hundreds of nodes). Also, maybe the simulated time scales are not enough for observing the conformational changes caused by the drugs.

Appendix A - Mathematical derivations

A.1 Mutual Information

This section is dedicated to prove the following relation:

$$M_{ij}(x,y) = H_i(x) + H_j(y) - H_{ij}(x,y) \tag{5.6}$$

In order to achieve that, we will make use of the definition of the joint probability distribution of two dependent variables:

$$\rho_{ij}(x,y) = \rho_i(x|y)\rho_j(y) \tag{5.7}$$

This relation is quite intuitive, and means that the joint probability of two events (for us that atom i has a fluctuation x and atom j has a fluctuation y), is the probability of having the first event given that the second one has already happened, multiplied by the probability of this second event. If the two events are independent, then $\rho_i(x|y) = \rho_i(x)$, as the probability of the first event is independent of the second one. If we insert this relation in the definition of the mutual information (5.3) and separate the logarithms:

$$M_{ij} = \int \rho_{ij}(x,y)[ln\rho_i(x|y) - ln\rho_i(x)]dxdy \tag{5.8}$$

We can integrate the variable y in the second term, obtaining $\int \rho_i ln(\rho_i(x)) = H_i(x)$. Inserting relation (5.7) in the first term:

$$\int \rho_{ij}(x,y)ln(\rho_i(x|y)) = \int \rho_j(y)\rho_i(x|y)ln(\rho_i(x|y))dxdy = -\int \rho_j(y)dyH_i(x|y) \tag{5.9}$$

where $\int \rho_j(y)dy = 1$, as the probability is normalized. Therefore we have obtained that:

$$M_{ij}(x,y) = H_i(x) - H_i(x|y) \tag{5.10}$$

In relation (5.4), it appears the joint entropy of the two variables, $H_{ij}(x,y)$, similarly, we are going to relate this magnitude with the conditional and marginal entropies, starting from the definition:

$$H_{ij}(x,y) = \int \rho_{ij}(x,y) ln(\rho_{ij}(x,y))dxdy = \int \rho_{ij}(x,y) ln(\rho_i(x|y)\rho_j(y))dxdy \tag{5.11}$$

In the last equality we have used again relation (5.7), separating the product inside the logarithm we have two terms $ln((\rho_i(x|y)\rho_j(y)) = ln((\rho_i(x|y)) + ln(\rho_j(y))$; where in the second we can integrate the variable x to obtain $H_j(y)$ and in the first one we can use again (5.7):

$$\int \rho_{ij}(x,y) ln\rho_i(x|y)dxdy = \int \rho_j(y)\rho_i(x|y) ln(\rho_i(x|y))dxdy = H_i(x|y) \tag{5.12}$$

Therefore, we have obtained the following relation:

$$H_{ij}(x,y) = H_j(y) + H_i(x|y) \Rightarrow H_i(x|y) = H_{ij}(x,y) - H_j(y) \tag{5.13}$$

Finally, if we substitute $H_i(x|y)$ from Eq. (5.13) in Eq. (5.10), we finally obtain that:

$$M_{ij}(x,y) = H_i(x) + H_j(y) - H_{ij}(x,y) \tag{5.14}$$

A.2 Courrent flow betweenness centrality

Previously, we have already defined our network as a matrix that has the information of the contacts and their mutual information, this will be the adjacency matrix of our graph A. The requirements for A is that the graph must be connected and the weights of the edges must be positive, this is our case as the mutual information and the contact map are positive magnitudes. From A we can define the Laplacian of the network:

$$L = D - A \tag{5.15}$$

Where D is the degree matrix, a diagonal matrix formed by summing the rows of A. The Laplacian of the graph has the information of the degree of the nodes in the diagonal. This tell us how well connected this vertex is, and also the connections that makes with their respective weights.

As stated before, one of the characteristics of the network analysis is that we have to define where the signal originates and where it ends before we start our calculations, this will be done mathematically using a supply vector, $b(v)$, which has value 1 if the vertex v is a source, -1 if it is a sink and 0 otherwise.

To introduce the current flow of the graph we have to provide our edges with an orientation, this is needed since, in the end, we want to account for the directionality of signals or perturbations within our network. One orientated edge will be denoted as $\vec{e}$, with coordinates $\vec{e} = (u, v)$, which mean that the edge $\vec{e}$ connects nodes u and v, its direction is set arbitrarily. The set of all oriented edges is called $\vec{E}$. The previous definition of the supply vector together with our adjagency matrix *univocally* determines the current flow of the graph $x(\vec{e})$. The current flow takes as input one orientated edge and gives a positive value if the current goes in the direction of $\vec{e}$ and negative if it goes against it, the value of $x(\vec{e})$ will have the information that describes how the signal is transmitted throughout our graph. This approach is based on electric circuits, the adjacency matrix has the information of the values of the electrical conductances connecting points in the circuit and the supply vector tells us where the current starts and ends, therefore it is clear that once we have our circuit defined with the electrical resistances and the electrical source (and sink) the electrical current that flows through the circuit is known, for us the current has the information of how easily a signal is transmitted in our protein, according to the properties that we have projected in our adjacency matrix.

The current flow is conserved throughout our graph aside from the sources or sinks and there is not any external current entering the circuit aside from the one defined in the supply vector, therefore Kirchhoff's laws will be followed, these are that the sum of the current of orientated nodes forming a cycle is zero and that the sum of the currents that flow into one node and the ones that flows out of that node is defined by the supply vector, mathematically:

$$\sum_{(v,w)\in\vec{E}} x(v,w) - \sum_{(u,v)\in\vec{E}} x(u,v) = b(v) \tag{5.16}$$

where the pairs (v, w) and (u, v) are the coordinates of oriented edges belonging $\vec{E}$, with $b(v) = 0$ for vertices not defined as sources or sinks. Continuing with the electrical jargon, we will use Ohm's law to define the potential difference of a vertex:

$$\hat{p}(\vec{e}) = \frac{x(\vec{e})}{c(e)} \tag{5.17}$$

where $c(e)$ are the weights of the edges [6] that appear in the adjacency matrix (and therefore in the Laplacian). We can come back to the non-directional point of view to define the *absolute potentials*, $p(v)$, a vector that gathers the potential differences of the nodes up to a common constant, with this definition we are projecting the directional information provided by the current flow in the absolute potentials, and we are not loosing any information since the current will flow from the higher potential values to the lower ones.

The potentials are very useful as they have the same information of the current flow, furthermore they connect the Laplacian to the supply vector:

$$Lp = b \tag{5.18}$$

To prove this relation, we will start with the left part of the equation, doing matrix product of the first row of the Laplacian with the potentials, which has the information of the first node:

$$\sum_{i=2}^{N} c(1,i)p(1) - c(1,2)p(2) - ... - c(1,N)p(N) \tag{5.19}$$

We recall that the sum is the degree of the node, which is the diagonal of the Laplacian, expanding the sum and grouping:

$$c(1,2)[p(1) - p(2)] + c(1,3)[p(1) - p(3)] + ... + c(1,N)[p(1) - p(N)] \tag{5.20}$$

where the potential difference $p(1) - p(2)$ is the previously called $\hat{p}(\vec{e}) = \hat{p}(1,2)$, using Ohm's law Eq. (5.20) results in the sum of all currents passing through node the first node[7], and applying Kirchhoff's law of currents we get the supply vector at this

[6] note that in the definition of the edge weights $c(e)$, e is also defined by two coordinates but it is not oriented and it does not have the vector sign, as this quantities do not have directionality.

[7] With the notation followed, all currents have positive sign but the value can be negative as it depends on the value of the difference $p(1) - p(2)$.

node, so equation (5.18) holds, before multiplying on the left by the inverse of the Laplacian to separate the potentials, we have to make a small comment about it, it is not possible as the laplatian of the graph will not have inverse; looking at the definition, the vector $V = (1\,1...1)^\top$ is an eigenvector with eigenvalue 0, this means that this set of equations:

$$p = L^{-1}b \tag{5.21}$$

will have infinite solutions, up to a constant, one approach is exploiting the fact that the laplatian accepts Moore–Penrose generalized inverse [123], but a simpler method consists in setting the potential of the first sink to zero, turning it into our potential reference, which is fine because we only care about potential differences. Mathematically this is done by defining a reduced Laplacian $\tilde{L}$, a matrix defined by setting the row and column of that node to zero, for example if that node is the first one we would get:

$$p = \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \tilde{L}^{-1} \end{pmatrix} \cdot b \tag{5.22}$$

Once we have the potentials the current flow betweeness of node i, τ_i, will be easily computed, as it is defined as the sum of all currents passing through i:

$$\tau_i = \frac{1}{2|S_0|} \sum_{s,v} |x_s(i,v)| \tag{5.23}$$

Where S_0 is the number of sources, the sum is done for all sources defined in our signal s and for all nodes v that are connected with the target node i. We have to take into account that each source will define a different supply vector $b_s(i)$ which will be also translated into a different current flow $x_s(\vec{e})$. The definition of the betweeness centrality can also be expressed in terms of the potentials as well using Ohm's law:

$$\tau_i = \frac{1}{2|S_0|} \sum_{s,j} |p_i(s) - p_j(s)| A_{ij} \tag{5.24}$$

which will be the equation used in practice, since the potentials can easily be computed from the Laplatian using Eq. (5.20), the notation $p_i(s)$ describes the potential of vertex i when the potentials are computed from the source s.

Appendix B - List of Abbreviations

Frequently used abbreviations

CaM Calmodulin

CaMBD Calmodulin Binding Domain

MD Molecular Dynamics

RMSD Root Mean Square Deviation

RMSF Root Mean Square Fluctuation

WT Wild Type

Chapter 1: Introduction

CaM Calmodulin

C terminus Carboxyl terminus

Cryo-EM Cryo Electron Microscopy

DNA deoxyribonucleic acid

MD Molecular Dynamics

NMR Nuclear Magnetic Resonance

N terminus amino terminus

RNA ribonucleic acid

RMSD Root Mean Square Deviation

RMSF Root Mean Square Fluctuation

Chapter 2

CaM Calmodulin

CaMBD Calmodulin Binding Domain

CRD Calcium Responsive Domain

CTD Carboxyl Terminal Domain

C lobe Carboxyl lobe of calmodulin (last in order by sequence)

FRET Förster Resonance Energy Transfer

MD Molecular Dynamics

N conformation Native conformation of the wild type, looking at the angle of the tryptophan with the rest of the helix A of KCNQ2

N lobe Amino lobe of calmodulin (first in order by sequence)

Tilted Tilted conformation of the wild type, looking at the angle of the tryptophan with the rest of the helix A of KCNQ2

VSD Voltage Sensor Domain

W344R-N Native like conformation of the W344R mutation, looking at the angle of the arginine with the rest of the helix A of KCNQ2

W344R-T Tilted like conformation of the W344R mutation, looking at the angle of the arginine with the rest of the helix A of KCNQ2

WHAM Weighted Histogram Analysis Method

WT Wild Type

Chapter 3

CaM Calmodulin

CaMBD Calmodulin Binding Domain

CaM C lobe Carboxyl lobe of calmodulin (first in order by sequence)

CRD Calcium Responsive Domain

MD Molecular Dynamics

REST2 Replica Exchange with Solute Scaling

SASA Solvent Accessible Surface Area

ViSD Voltage inSensitive Domain

Chapter 4

CRD Calcium Responsive Domain

Chapter 5

CaM C lobe Carboxyl lobe of calmodulin (last in order by sequence)

CaM N lobe Amino lobe of calmodulin (first in order by sequence)

CFB Current Flow Betweenness centrality

CRD Calcium Responsive Domain

CTD Carboxyl Terminal Domain

RTG Retigabine

RMSD Root Mean Square Deviation

RMSF Root Mean Square Fluctuation

ZTZ ZTZ240

Appendix C - Supplementary information guide

CH.2 - Rosetta calculations - 344 mutants

Rosetta calculations of the binding energy of several variants of the $K_v7.2$ channel, these calculations were performed with the package called flexddG [55] of the rosetta software (https://www.rosettacommons.org/), more info also in the tutorial of the package [124].

We have all calculations separated in folders named after the mutation under study, these are W344E, W344F, W344G, W344Q and W344R.

All folders containing the outputs have the same architecture, which will be described now:

inputs/6FEG/: Inputs needed by rosetta

→ '6FEG.pdb': input structure for the protocol, from [41].

→ 'chains_to_move.txt': chain moved.

→ 'mutation.resfile': definition of the analyzed mutation.

rosetta_scripts/: Code of the protocol with its options.

→ 'flexddG.wt.xml': main code for flexddG protocol.

→ 'flexddG.options': parser options needed for rosetta command.

output/6FEG/01/: output of the first simulation, up to output/6FEG/50/, output data bases or structures are not provided due to storage limitation.

→ 'rosetta.out': main output file, the scores obtained for the structures are appended to this file.

analysis_output/: output in spreadsheet format.

→ '-results.csv': $\Delta\Delta G$ and score terms of the backrub steps.

→ '-struct_scores_results.csv': same as results.csv but with different order

./: parent folder of the simulation with extra files.

→ 'run.py': python script that performs the whole simulation, its output is the final $\Delta\Delta G$

→ 'extract_structures.py': reads output databases and extract structures in pdb format

→ 'analyze_flex_ddG.py': reads the output and writes the previously commented spreadsheet output files

CH.2 - Simulations of CaM C lobe and helix hA of K_v7.2 channel

Molecular dynamics simulations of CaM C lobe bounded to helix hA were performed to understand the energetic interactions that would arise from the binding of the Native (W344R-N) and Tilted (W344R-N) configurations, each situation was simulated following the same workflow and can be respectively found in the folders: Mut-n, Mut-t and WT

./: parent folder of the simulation with coordinate (.pdb) and topology files (.psf) with the most probable configuration of the run.

step1_eminth_nvt/: Minimization and NVT thermalization

→ 'eqMT01.conf': first half of the first step of thermalization.

→ 'eqMT02.conf': second half of the first step of thermalization.

→ 'eqMT02.restart.coor': final coordinates of the first step of thermalization.

step2_th_npt/: NPT thermalization

→ 'eqMT03.conf': first half of the second step of thermalization.

→ 'eqMT04.conf': second half of the second step of thermalization.

→ 'eqMT04.restart.coor': final coordinates of the second step of thermalization.

step3_prod/: Production

→ 'md.conf': input run for the production run.

→ 'md.restart.coor': binary file with the end of the run.

→ 'angles.png': time series of the angle of the run, angle definition can be found on the main text, chapter 3, Fig. 3.3.

→ 'rmsf.png': root mean square fluctuation of each amino acid on the sequence.

→ 'rms.png': total root mean square deviation of the sequence.

CH.2 - Simulations of K_v7.2 CRD

Molecular dynamics simulations of helices hA, TW, and hB were performed for three different configurations, wild type, introducing the mutation W344R and WT but starting from a tilted conformation; each situation was simulated following the same workflow and can be respectively found in the folders: MUT WT WT_tilted

***./*:** parent folder of the simulation with coordinate (.pdb) and topology files (.psf).

step1_min_nvt: Minimization of the input structure followed by a NVT thermalization.

→ 'emint1_hAB_mut.conf': input file for the minimization and first step of thermalization.

→ 'emint1_hAB_mut.restart.coor': binary file with the coordinates at the end of the first thermalization step.

step2_npt: NPT thermalization.

→ 'th2_hAB_mut_NPT.conf': input files of the second step of thermalization.

→ 'th2_hAB_mut_NPT.restart.coor': binary file with the coordinates at the end of the second thermalization step.

step3_prod: Production. This last step was repeated three times to obtain three different replicas, each of 100 ns of duration.

→ 'hAB_mut_long_run.conf': input run for the production run.

→ **replica_1**: results of the first production replica.

⋆ 'angles.png': time series of the angle of the run, angle definition can be found on the main text, chapter 3, Fig. 3.3.

⋆ 'rmsf.png': root mean square fluctuation of each amino acid on the sequence.

⋆ 'rms.png': total root mean square deviation of the sequence.

⋆ 'hAB_mut_long_run.restart.coor': binary file with the coordinates at the end of the production.

→ **replica_2**: results of the second production replica, same as replica 1.

→ **replica_3**: results of the third production replica, same as replica 1.

CH.2 - Rosetta calculations of double mutants

Supporting information of the calculations of the binding energy of several variants of the $K_V7.2$ channel, these calculations were performed with the package called flexddG [55] of the rosetta software (https://www.rosettacommons.org/), more info also in the tutorial of the package [124].

We have all calculations separated in folders named after the mutation under study, since all calculations have the mutation W344R, suffix MUT is added: Q341A-MUT, Q341C-MUT, Q341D-MUT, Q341F-MUT, Q341H-MUT, Q341K-MUT, Q341M-MUT, Q341C-MUT, Q341E-MUT, Q341G-MUT, Q341N-MUT, Q341I-MUT and Q341L-MUT.
All folders containing the outputs have the same architecture, which will be described now:

inputs/6FEG/: Inputs needed by rosetta

→ '6FEG.pdb': input structure for the protocol, from [41].

→ 'chains_to_move.txt': chain moved.

→ 'mutation.resfile': definition of the analyzed mutation.

rosetta_scripts/: Code of the protocol with its options.

→ 'flexddG.wt.xml': main code for flexddG protocol.

→ 'flexddG.options': parser options needed for rosetta command.

output/6FEG/01/: output of the first simulation, up to output/6FEG/50/, output data bases or structures are not provided due to storage limitation.

→ 'rosetta.out': main output file, the scores obtained for the structures are appended to this file.

analysis_output/: output in spreadsheet format.

→ '-results.csv': $\Delta\Delta G$ and score terms of the backrub steps.

→ '-struct_scores_results.csv': same as results.csv but with different order

./: parent folder of the simulation with extra files.

→ 'run.py': python script that performs the whole simulation, its output is the final $\Delta\Delta G$

→ 'extract_structures.py': reads output databases and extract structures in pdb format

→ 'analyze_flex_ddG.py': reads the output and writes the previously commented spreadsheet output files

CH.2 - Simulations of double mutants

Molecular dynamics simulations of helices hA, TW, and hB were performed for several double mutant sequences, each situation was simulated following the same workflow and can be respectively found in the folders:
Q341A-W344R_double_mutant, Q341L-W344R_double_mutant,
Q341P-W344R_double_mutant, Q341C-W344R_double_mutant,
Q341M-W344R_double_mutant, Q341V-W344R_double_mutant,
Q341I-W344R_double_mutant, Q341N-W344R_double_mutant and W344R_single_mutant.

The folders contain the same information for each simulation, which is the following:

./: parent folder of the simulation with coordinate (.pdb) and topology files (.psf).

step1_min_nvt: Minimization and NVT thermalization.

→ 'emint1_341_mut_NVT_base.conf': input file for the minimization and first step of thermalization.

step2_npt: NPT thermalization.

→ 'emint1_341_mut_NPT_base.conf': input files for the second step of thermalization.

step3_prod: Production, this last step was repeated three times to obtain three different replicas that make up to $\sim 450ns$ of production time.

→ 'hAB_mut_long_run.conf': input run for the production run.

→ **replica_1**: results of the first production replica.

⋆ 'angles.png': time series of the angle of the run, angle definition can be found on the main text, chapter 3, Fig. 3.3.

⋆ 'rmsf.png': root mean square fluctuation of each amino acid on the sequence.

⋆ 'rms.png': total root mean square deviation of the sequence.

→ **replica_2**: results of the second production replica, same as replica 1.

→ **replica_3**: results of the third production replica, same as replica 1.

CH.3 - REST2 simulations

Replica exchange with solute scaling simulations were performed to observe the folding of a small fragment of SK2, under different conditions:
Immersed in water alone.

In presence of CaM C-lobe starting from the "one turn" configuration.

In presence of CaM C-lobe starting from the "two turns" configuration.

In presence of CaM C-lobe starting from the "one turn" configuration with the Trp outward.

In presence of CaM C-lobe starting from the "two turns" configuration with the Trp inward.

And also some IQ motifs were simulated immersed in water.
All these simulations contains the same supplementary information:

***Min/*:** Minimization.

→ 'min.mdp': input parameters for the minimization.

→ 'min.gro': output structure of the minimization.

***NVT/*:** Then NVT equilibration.

→ 'nvt.mdp': input parameters for the NVT equilibration.

→ 'nvt.gro': output structure of the NVT equilibration

→ 'processed.top': processed topology for equilibration, atom names of atoms comprising hot subsystem have _ attached for scaling with plumed script.

***NPT/*:** NPT equilibration.

→ 'npt.mdp': input parameters for the NPT equilibration.

→ 'npt.gro': output structure of the NPT equilibration

→ 'processed.top': processed topology for equilibration, atom names of atoms comprising hot subsystem have _ attached for scaling with plumed script.

***Prod/*:** Production runs.

→ 'md.mdp': input parameters for the production.

→ 'md.gro': final snapshot of the production run.

→ 'processed.top': processed topology for production, atom names of atoms comprising hot subsystem have _ attached for scaling with plumed script.

→ 'exch_0.png': exchanges of the first replica, following its path along the replicas through time.

→ 'exch_all.png': exchanges of all replicas through time.

→ 'exchange_count.xvg': number of exchanges between the replicas and its probability.

***./*:** in the parent folder the script for making the different topologies is given, named "gentopHREX.sh", it needs a running gromacs compilation patched with plumed, effective temperature range can be selected in the script, it is run with the command bash gentopHREX.sh and generates the scaled topologies in different folders R0...R${N} where N is the number of replicas minus one.

CH.3 - Metadynamics simulations

Two see if there was a barrier separating the conformations of the angle formed by the triptophan with CaM, well-tempered metadynamics simulations [84] were carried out, there was no need for thermalize the systems, since the previously equilibrated 'one turn' and 'two turns' systems were used. Results for the "One turn" and "Two turns" can be found in the folders OneTurn and TwoTurns, respectively, in both cases the same information is given:

→ 'md.mdp': input parameters for the molecular dynamics.

→ 'md.gro': input coordinate file for the simulation.

→ 'topol.top': compressed topology of the system.

→ 'metadynamics.png': time series of the angles of the whole simulation and free energy surfaces obtained for the last 200ns.

→ 'plumed.dat': input parameters for the metadynamics, in which the collective variable, parameters for the added Gaussian functions and the funnel potential are defined.

CH.4 - REST2 simulations in aqueous solution

To observe the impact of the W344R mutation on the folding of a small fragment of the KCNQ2 channel when immersed in water, we performed replica exchange with solute scaling (REST2) simulations, separated in two folders: WT and W344R. Inside of each folder some input and output files of the whole protocol can be found:

***Min/*:** Each replica was energy-minimized.

→ 'min.mdp': input parameters for the minimization.

→ 'min.gro': output structure of the minimization.

***NVT/*:** Then each replica was equilibrated at NVT ensemble.

→ 'nvt.mdp': input parameters for the NVT equilibration.

→ 'nvt.gro': output structure of the NVT equilibration

→ 'processed.top': processed topology for equilibration, atom names of atoms comprising hot subsystem have _ attached for scaling with plumed script.

***NPT/*:** Then each replica was equilibrated at NPT ensemble.

→ 'npt.mdp': input parameters for the NPT equilibration.

→ 'npt.gro': output structure of the NPT equilibration

→ 'processed.top': processed topology for equilibration, atom names of atoms comprising hot subsystem have _ attached for scaling with plumed script.

***Prod/*:** Production runs.

→ 'md.mdp': input parameters for the production.

→ 'md.gro': final snapshot of the production run.

→ 'processed.top': processed topology for production, atom names of atoms comprising hot subsystem have _ attached for scaling with plumed script.

→ 'exch_0.png': exchanges of the first replica, following its path along the replicas through time.

→ 'exch_all.png': exchanges of all replicas through time.

→ 'exchange_count.xvg': number of exchanges between the replicas and its probability.

***./*:** in the parent folder the script for making the different topologies is given, named "gentopHREX.sh", it needs a running gromacs compilation patched with plumed, effective temperature range can be selected in the script, it is run with the command bash gentopHREX.sh and generates the scaled topologies in different folders R0...R${N} where N is the number of replicas minus one.

CH.4 - REST2 cotranslational simulations

To observe the cotranslational folding of the small fragment of the KCNQ2 channel, replica exchange with solute scaling (REST2) simulations were performed. Wild type and mutation W344R were simulated following the same workflow and can be respectively found in the folders: WT and W344R. Inside of each folder some input and output files of the whole protocol can be found:

***Min/*:** → 'min.mdp': input parameters for the minimization.

→ 'min.gro': output structure of the minimization.

→ 'topol_min.top': topology used for minimization (without position restraints.

***therm/*:** → 'nvt.mdp': input parameters for the NVT equilibration.

→ 'nvt.gro': output structure of the NVT equilibration

→ 'npt.mdp': input parameters for the NPT equilibration.

→ 'npt.gro': output structure of the NPT equilibration

→ 'topol_therm.top': processed topology for equilibration.

→ 'rms.png': RMSD of nascent chain and ribosome during equilibration steps.

***Prod/*:** Production runs.

→ 'prod.mdp': input parameters for the production.

→ 'processed_prod.top': processed topology for production, atom names of atoms comprising hot subsystem have _ attached for scaling with plumed script.

→ **replica1**: results of the first production replica.

⋆ 'exch_0.png': exchanges of the first replica, following its path along the replicas through time.

⋆ 'exch_all.png': exchanges of all replicas through time.

⋆ 'exch_count.xvg': statistics of the exchanges.

⋆ 'rms.png': total root mean square deviation of the sequence.

⋆ 'rmsf.png': total root mean square deviation of the sequence.

→ **replica2**: results of the second production replica.

→ **replica3**: results of the third production replica.

CH.5 - Full channel simulations

Molecular dynamics simulation of the whole channel structures of Kv7.2 and SK4 were performed, in the presence and in the absence of some drugs, whose supplementary information is stored in: Kv7.2_Apo, Kv7.2_Retigabine, Kv7.2_Ztz240, SK4_APO and SK4_Riluzole.

In the case of the SK4_Riluzole folder, also docking parameters are provided in the file docking_parameters.pdf. Aside of that, each folder contains the following information:

***./*:** Parent folder.

→ 'topol.top': GROMACS topology file.

→ 'docking_parameters.pdf': parameters of the docking calculations to place riluzole in its binding pocket, in the case of SK4_Riluzole folder.

***toppar/*:** topology and parameter files that are required by the topology file.

***Therm/*:** 6 steps of NPT thermalization, with constraints applied to the membrane and protein, these constraints are gradually decreased in each step.

→ 'step6.0_minimization.mdp': input parameters for the minimization.

→ 'step6.1_equilibration.mdp'... 'step6.6_equilibration.mdp': input parameter files for the equilibration steps.

→ 'step6.6_equilibration.gro': final structure after thermalization, will be the input for production.

→ 'rms_therm.png': root mean square deviation of the protein atoms during the last step of thermalization.

***prod/*:** 500 ns long production runs at NPT ensemble.

→ 'step7_production.mdp': input parameters for the production runs.

→ '**replica1/**': output of the first production replica

⋆ '${NAME}-1.gro.xtc': final coordinates of the run, in binary gromacs .xtc file

⋆ 'rmsd-${NAME}-1.png': root mean square of the protein atoms, separated into calmoduline, transmembrane domain and calmodulin binding domains.

⋆ 'rmsf${NAME}-1.png': root mean square fluctuations of the amino acids of the sequence of the channel.

→ '**replica2/**': output of the second production replica, same content as replica1.

→ '**replica3/**': output of the third production replica, same content as replica1.

Bibliography

[1] Janire Urrutia et al. “An epilepsy-causing mutation leads to co-translational misfolding of the Kv7.2 channel”. eng. In: *BMC biology* 19.1 (May 2021). 34020651[pmid], pp. 109–109. ISSN: 1741-7007. DOI: 10.1186/s12915-021-01040-1.

[2] Ralph Wissmann et al. “A Helical Region in the C Terminus of Small-Conductance Ca^{2+}-Activated K^+ Channels Controls Assembly with apo-Calmodulin”. In: *J. Biol. Chem.* 277 (2002), pp. 4558–4564. DOI: 10.1074/jbc.M109240200.

[3] Lily T.-Y. Cho et al. In: *Structure* 26.4 (Apr. 2018), 533–544.e3. ISSN: 0969-2126.

[4] Lucie Delemotte et al. “Intermediate states of the Kv1.2 voltage sensor from atomistic molecular dynamics simulations”. In: *Proceedings of the National Academy of Sciences* 108.15 (2011), pp. 6109–6114. ISSN: 0027-8424. DOI: 10.1073/pnas.1102724108.

[5] Morten Ø. Jensen et al. “Mechanism of Voltage Gating in Potassium Channels”. In: *Science* 336.6078 (2012), pp. 229–233. DOI: 10.1126/science.1216533.

[6] Aashish Manglik et al. “Structure-based discovery of opioid analgesics with reduced side effects”. In: *Nature* 537.7619 (Sept. 2016), pp. 185–190. ISSN: 1476-4687. DOI: 10.1038/nature19112.

[7] Aaron Chevalier et al. “Massively parallel de novo protein design for targeted therapeutics”. In: *Nature* 550.7674 (Oct. 2017), pp. 74–79. ISSN: 1476-4687. DOI: 10.1038/nature23912.

[8] K Wüthrich. “Protein structure determination in solution by NMR spectroscopy.” In: *Journal of Biological Chemistry* 265.36 (1990), pp. 22059–22062. ISSN: 0021-9258. DOI: https://doi.org/10.1016/S0021-9258(18)45665-7.

[9] Isabel Moraes et al. “Membrane protein structure determination — The next generation”. In: *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1838.1, Part A (2014). Structural and biophysical characterisation of membrane protein-ligand binding, pp. 78–87. ISSN: 0005-2736. DOI: https://doi.org/10.1016/j.bbamem.2013.07.010.

[10] Edward H. Egelman. “The Current Revolution in Cryo-EM”. In: *Biophysical Journal* 110.5 (Mar. 2016), pp. 1008–1012. ISSN: 0006-3495. DOI: 10.1016/j.bpj.2016.02.001.

[11] Yukihiko Sugita et al. “Cryo-EM structure of the Ebola virus nucleoprotein–RNA complex at 3.6 Å resolution”. In: *Nature* 563.7729 (Nov. 2018), pp. 137–140. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0630-0.

[12] Heena Khatter et al. “Structure of the human 80S ribosome”. In: *Nature* 520.7549 (Apr. 2015), pp. 640–645. ISSN: 1476-4687. DOI: 10.1038/nature14427.

[13] Ji Sun and Roderick MacKinnon. “Structural Basis of Human KCNQ1 Modulation and Gating”. In: *Cell* 180.2 (2020), 340–347.e9. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2019.12.003.

[14] Scott A. Hollingsworth and Ron O. Dror. “Molecular Dynamics Simulation for All”. In: *Neuron* 99.6 (2018), pp. 1129–1143. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2018.08.011.

[15] James C. Phillips et al. “Scalable molecular dynamics on CPU and GPU architectures with NAMD”. In: *The Journal of Chemical Physics* 153.4 (2020), p. 044130. DOI: 10.1063/5.0014475.

[16] Mark James Abraham et al. “GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers”. In: *SoftwareX* 1-2 (2015), pp. 19–25. ISSN: 2352-7110. DOI: https://doi.org/10.1016/j.softx.2015.06.001. URL: https://www.sciencedirect.com/science/article/pii/S2352711015000059.

[17] Daan Frenkel and Berend Smit. “Chapter 4 - Molecular Dynamics Simulations”. In: *Understanding Molecular Simulation (Second Edition)*. Ed. by Daan Frenkel and Berend Smit. Second Edition. San Diego: Academic Press, 2002, pp. 63–107. ISBN: 978-0-12-267351-1. DOI: https://doi.org/10.1016/B978-012267351-1/50006-7.

[18] Jorge Kohanoff. *Electronic Structure Calculations for Solids and Molecules: Theory and Computational Methods*. Cambridge University Press, 2006. DOI: 10.1017/CBO9780511755613.

[19] Chuan Tian et al. "ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution". In: *Journal of Chemical Theory and Computation* 16.1 (Jan. 2020), pp. 528–552. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.9b00591.

[20] Leela S. Dodda et al. "LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands". In: *Nucleic Acids Research* 45.W1 (Apr. 2017), W331–W336. ISSN: 0305-1048. DOI: 10.1093/nar/gkx312.

[21] Jing Huang and Alexander D MacKerell Jr. "CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data". eng. In: *Journal of computational chemistry* 34.25 (Sept. 2013). 23832629[pmid], pp. 2135–2145. ISSN: 1096-987X. DOI: 10.1002/jcc.23354.

[22] Alexander D. Mackerell Jr. "Empirical force fields for biological macromolecules: Overview and issues". In: *Journal of Computational Chemistry* 25.13 (2004), pp. 1584–1604. DOI: https://doi.org/10.1002/jcc.20082.

[23] Paul Bauer, Berk Hess, and Erik Lindahl. *GROMACS 2022 Manual*. Version 2022. Feb. 2022. DOI: 10.5281/zenodo.6103568.

[24] James C. Phillips et al. "Scalable molecular dynamics with NAMD". In: *Journal of Computational Chemistry* 26.16 (2005), pp. 1781–1802. DOI: https://doi.org/10.1002/jcc.20289.

[25] Darryl K. Granner Peter A. Mayes Victor W. Rodwell Robert K. Murray. *Harper's Illustrated Biochemistry*. 26th. McGraw-Hill Medical, 2003.

[26] H. Lodish et al. *Molecular Cell Biology*. 4th. New York: W. H. Freeman, 2000.

[27] Marija Liutkute, Ekaterina Samatova, and Marina V. Rodnina. "Cotranslational Folding of Proteins on the Ribosome". In: *Biomolecules* 10.1 (2020). ISSN: 2218-273X. DOI: 10.3390/biom10010097.

[28] Michael Thommen, Wolf Holtkamp, and Marina V Rodnina. "Co-translational protein folding: progress and methods". In: *Current Opinion in Structural Biology* 42 (2017). Folding and binding • Proteins: Bridging theory and experiment, pp. 83–89. ISSN: 0959-440X. DOI: https://doi.org/10.1016/j.sbi.2016.11.020.

[29] Ameer N. Thompson et al. “Mechanism of potassium-channel selectivity revealed by Na+ and Li+ binding sites within the KcsA pore”. In: *Nature Structural & Molecular Biology* 16.12 (2009), pp. 1317–1324. ISSN: 1545-9985. DOI: 10.1038/nsmb.1703. URL: https://doi.org/10.1038/nsmb.1703.

[30] Eider Núñez, Arantza Muguruza-Montero, and Alvaro Villarroel. “Atomistic Insights of Calmodulin Gating of Complete Ion Channels”. In: *International Journal of Molecular Sciences* 21.4 (2020). ISSN: 1422-0067. DOI: 10.3390/ijms21041285.

[31] Carolina Gomis-Pérez et al. “Homomeric Kv7.2 current suppression is a common feature in KCNQ2 epileptic encephalopathy”. In: *Epilepsia* 60.1 (2019), pp. 139–148. DOI: https://doi.org/10.1111/epi.14609.

[32] Snezana Maljevic and Holger Lerche. “Chapter 2 - Potassium channel genes and benign familial neonatal epilepsy”. In: *Genetics of Epilepsy*. Ed. by Ortrud K. Steinlein. Vol. 213. Progress in Brain Research. Elsevier, 2014, pp. 17–53. DOI: https://doi.org/10.1016/B978-0-444-63326-2.00002-8.

[33] William Humphrey, Andrew Dalke, and Klaus Schulten. “VMD – Visual Molecular Dynamics”. In: *Journal of Molecular Graphics* 14 (1996), pp. 33–38.

[34] Xiaoxiao Li et al. “Molecular basis for ligand activation of the human KCNQ2 channel”. In: *Cell Research* 31.1 (Jan. 2021), pp. 52–61. ISSN: 1748-7838. DOI: 10.1038/s41422-020-00410-8.

[35] Paolo Ambrosino et al. “Epilepsy-causing mutations in Kv7.2 C-terminus affect binding and functional modulation by calmodulin”. In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1852.9 (2015), pp. 1856–1866. ISSN: 0925-4439. DOI: https://doi.org/10.1016/j.bbadis.2015.06.012.

[36] Maria Virginia Soldovieri et al. “Novel KCNQ2 and KCNQ3 Mutations in a Large Cohort of Families with Benign Neonatal Epilepsy: First Evidence for an Altered Channel Regulation by Syntaxin-1A”. In: *Human Mutation* 35.3 (2014), pp. 356–367. DOI: https://doi.org/10.1002/humu.22500.

[37] Arantza Muguruza-Montero et al. “Do calmodulin binding IQ motifs have built-in capping domains?” In: *Protein Science* 30.10 (2021), pp. 2029–2041. DOI: https://doi.org/10.1002/pro.4170.

[38] Alessandro Alaimo et al. "Calmodulin Activation Limits the Rate of KCNQ2 K^+ Channel Exit from the Endoplasmic Reticulum *". In: *Journal of Biological Chemistry* 284.31 (July 2009), pp. 20668–20675. ISSN: 0021-9258. DOI: 10.1074/jbc.M109.019539.

[39] Guillaume Gotthard et al. "Chromophore Isomer Stabilization Is Critical to the Efficient Fluorescence of Cyan Fluorescent Proteins". In: *Biochemistry* 56.49 (Dec. 2017), pp. 6418–6422. ISSN: 0006-2960. DOI: 10.1021/acs.biochem.7b01088.

[40] Corinna Ottmann et al. "Applicability of superfolder YFP bimolecular fluorescence complementation in vitro". In: 390.1 (2009), pp. 81–90. DOI: doi:10.1515/BC.2009.008.

[41] Ganeko Bernardo-Seisdedos et al. "Structural basis and energy landscape for the Ca^{2+} gating and calmodulation of the Kv7.2 K^+ channel". In: *Proceedings of the National Academy of Sciences* 115.10 (2018), pp. 2395–2400. DOI: 10.1073/pnas.1800235115.

[42] Nurzian Ismail et al. "A biphasic pulling force acts on transmembrane helices during translocon-mediated membrane integration". eng. In: *Nature structural & molecular biology* 19.10 (Oct. 2012). 23001004[pmid], pp. 1018–1022. ISSN: 1545-9985. DOI: 10.1038/nsmb.2376.

[43] Ola B. Nilsson et al. "Cotranslational Protein Folding inside the Ribosome Exit Tunnel". In: *Cell Reports* 12.10 (Sept. 2015), pp. 1533–1540. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2015.07.065.

[44] José Nelson Onuchic, Zaida Luthey-Schulten, and Peter G. Wolynes. "THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective". In: *Annual Review of Physical Chemistry* 48.1 (1997). PMID: 9348663, pp. 545–600. DOI: 10.1146/annurev.physchem.48.1.545.

[45] Hahnbeom Park et al. "Automatic structure prediction of oligomeric assemblies using Robetta in CASP12". eng. In: *Proteins* 86 Suppl 1.Suppl 1 (Mar. 2018). 28913931[pmid], pp. 283–291. ISSN: 1097-0134. DOI: 10.1002/prot.25387.

[46] Nicholas A. Marze et al. "Efficient flexible backbone protein-protein docking for challenging targets". eng. In: *Bioinformatics (Oxford, England)* 34.20 (Oct. 2018). 29718115[pmid], pp. 3461–3469. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bty355.

[47] Jens Meiler and David Baker. "ROSETTALIGAND: Protein–small molecule docking with full side-chain flexibility". In: *Proteins: Structure, Function, and Bioinformatics* 65.3 (2006), pp. 538–548. DOI: https://doi.org/10.1002/prot.21086.

[48] Kalli Kappel and Rhiju Das. "Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Folding and Docking". In: *Structure* 27.1 (2019), 140–151.e5. ISSN: 0969-2126. DOI: https://doi.org/10.1016/j.str.2018.10.001.

[49] Brian D. Weitzner et al. "Modeling and docking of antibody structures with Rosetta". In: *Nature Protocols* 12.2 (Feb. 2017), pp. 401–416. ISSN: 1750-2799. DOI: 10.1038/nprot.2016.180.

[50] Rebecca F. Alford et al. "An Integrated Framework Advancing Membrane Protein Modeling and Design". In: *PLOS Computational Biology* 11.9 (Sept. 2015), pp. 1–23. DOI: 10.1371/journal.pcbi.1004398.

[51] Valerie Vaissier Welborn and Teresa Head-Gordon. "Computational Design of Synthetic Enzymes". In: *Chemical Reviews* 119.11 (June 2019), pp. 6613–6630. ISSN: 0009-2665. DOI: 10.1021/acs.chemrev.8b00399.

[52] Julia Koehler Leman et al. "Macromolecular modeling and design in Rosetta: recent methods and frameworks". In: *Nature Methods* 17.7 (July 2020), pp. 665–680. ISSN: 1548-7105. DOI: 10.1038/s41592-020-0848-2.

[53] Nicholas Metropolis et al. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092. DOI: https://doi.org/10.1063/1.1699114.

[54] Andrew Leaver-Fay et al. "Scientific benchmarks for guiding macromolecular energy function improvement". eng. In: *Methods in enzymology* 523 (2013). 23422428[pmid], pp. 109–143. ISSN: 1557-7988. DOI: 10.1016/B978-0-12-394292-0.00006-0.

[55] Kyle A. Barlow et al. "Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-Protein Binding Affinity upon Mutation". eng. In: *The journal of physical chemistry. B* 122.21 (May 2018). 29401388[pmid], pp. 5389–5399. ISSN: 1520-5207. DOI: 10.1021/acs.jpcb.7b11367.

[56] Rebecca F. Alford et al. "The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design". eng. In: *Journal of chemical theory and computation* 13.6 (June 2017). 28430426[pmid], pp. 3031–3048. ISSN: 1549-9626. DOI: 10.1021/acs.jctc.7b00125.

[57] Georg Kuenze et al. "Upgraded molecular models of the human KCNQ1 potassium channel". In: *PLOS ONE* 14.9 (Sept. 2019), pp. 1–33. DOI: 10.1371/journal.pone.0220415.

[58] Jiaren Zhang et al. "Identifying mutation hotspots reveals pathogenetic mechanisms of KCNQ2 epileptic encephalopathy". In: *Scientific Reports* 10.1 (Mar. 2020), p. 4756. ISSN: 2045-2322. DOI: 10.1038/s41598-020-61697-6.

[59] Jainab Khatun, Sagar D. Khare, and Nikolay V. Dokholyan. ""Can contact potentials reliably predict stability of proteins?"". In: *Journal of molecular biology* 336.5 (2004), pp. 1223–1238. DOI: https://doi.org/10.1016/j.jmb.2004.01.002.

[60] https://www.rosettacommons.org/docs/latest/rosetta_basics/Units-in-Rosetta. Accessed: April of 2022.

[61] https://github.com/ORBallesteros/Thesis/tree/chapter2/rosetta_calculations/344_mutants. Accessed: September of 2022.

[62] Pekka Mark and Lennart Nilsson. "Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K". In: *The Journal of Physical Chemistry A* 105.43 (Nov. 2001), pp. 9954–9960. ISSN: 1089-5639. DOI: 10.1021/jp003020w.

[63] Scott E. Feller et al. "Constant pressure molecular dynamics simulation: The Langevin piston method". In: *The Journal of Chemical Physics* 103.11 (1995), pp. 4613–4621. DOI: 10.1063/1.470648.

[64] Axel Brünger. "X-PLOR Version 3.1. A System for X-ray Crystallography and NMR". In: Yale University Press, 1993. ISBN: 978-0300054026.

[65] https://github.com/ORBallesteros/Thesis/tree/chapter2/CaM-Clobe_hA. Accessed: September of 2022.

[66] https://github.com/ORBallesteros/Thesis/tree/chapter2/channel_only_CRD. Accessed: September of 2022.

[67] Shankar Kumar et al. "THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method". In: *Journal of Computational Chemistry* 13.8 (1992), pp. 1011–1021. DOI: https://doi.org/10.1002/jcc.540130812.

[68] Grossfield,Alan,\OT1\textquotedblleftWHAM:theweightedhistogramanalysismethod\OT1\textquotedblright,version2.0.11,http://membrane.urmc.rochester.edu/wordpress/?page_id=126.

[69] https://github.com/ORBallesteros/Thesis/tree/chapter2/rosetta_calculations/double_mutants. Accessed: September of 2022.

[70] https://github.com/ORBallesteros/Thesis/tree/chapter2/double_mutants_CRD. Accessed: September of 2022.

[71] Rafael Ramis et al. "Molecular dynamics simulations of the calmodulin-induced α-helix in the SK2 calcium-gated potassium ion channel". In: *Journal of Biological Chemistry* 299.2 (2023). ISSN: 0021-9258. DOI: 10.1016/j.jbc.2022.102850. URL: https://doi.org/10.1016/j.jbc.2022.102850.

[72] Janire Urrutia et al. "The Crossroad of Ion Channels and Calmodulin in Disease". In: *International Journal of Molecular Sciences* 20.2 (2019). ISSN: 1422-0067. DOI: 10.3390/ijms20020400.

[73] Allen R. Rhoads and Felix Friedberg. "Sequence motifs for calmodulin recognition". In: *The FASEB Journal* 11.5 (1997), pp. 331–340. DOI: https://doi.org/10.1096/fasebj.11.5.9141499.

[74] Nicholas J. Anthis, Michaeleen Doucleff, and G. Marius Clore. "Transient, Sparsely Populated Compact States of Apo and Calcium-Loaded Calmodulin Probed by Paramagnetic Relaxation Enhancement: Interplay of Conformational Selection and Induced Fit". In: *J. Am. Chem. Soc.* 133.46 (Nov. 2011), pp. 18966–18974. DOI: 10.1021/ja2082813.

[75] Fei Liu et al. "Molecular Mechanism of Multispecific Recognition of Calmodulin through Conformational Changes". In: *Proc. Natl. Acad. Sci. U. S. A.* 114.20 (2017), E3927–E3934. ISSN: 0027-8424. DOI: 10.1073/pnas.1615949114.

[76] Chia-Hsueh Lee and Roderick MacKinnon. "Activation mechanism of a human SK-calmodulin channel complex elucidated by cryo-EM structures". In: *Science* 360.6388 (2018), pp. 508–513. DOI: 10.1126/science.aas9466.

[77] Andrew Waterhouse et al. "SWISS-MODEL: homology modelling of protein structures and complexes". In: *Nucleic acids research* 46.W1 (July 2018). 29788355[pmid], W296–W303. ISSN: 1362-4962. DOI: 10.1093/nar/gky427.

[78] Stefan Bienert et al. "The SWISS-MODEL Repository-new features and functionality". In: *Nucleic acids research* 45.D1 (Jan. 2017). 27899672[pmid], pp. D313–D319. ISSN: 1362-4962. DOI: 10.1093/nar/gkw1132.

[79] Rafael C. Bernardi, Marcelo C.R. Melo, and Klaus Schulten. "Enhanced sampling techniques in molecular dynamics simulations of biological systems". In: *Biochimica et Biophysica Acta (BBA) - General Subjects* 1850.5 (2015). Recent developments of molecular dynamics, pp. 872–877. ISSN: 0304-4165. DOI: https://doi.org/10.1016/j.bbagen.2014.10.019.

[80] Lingle Wang, Richard A. Friesner, and B. J. Berne. "Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2)". In: *The Journal of Physical Chemistry B* 115.30 (Aug. 2011), pp. 9431–9438. ISSN: 1520-6106. DOI: 10.1021/jp204407d.

[81] Alessandro Laio and Michele Parrinello. "Escaping free-energy minima". In: *Proceedings of the National Academy of Sciences* 99.20 (2002), pp. 12562–12566. DOI: 10.1073/pnas.202427399.

[82] Yuji Sugita and Yuko Okamoto. "Replica-exchange molecular dynamics method for protein folding". In: *Chemical Physics Letters* 314.1 (1999), pp. 141–151. ISSN: 0009-2614. DOI: https://doi.org/10.1016/S0009-2614(99)01123-9.

[83] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. "Metadynamics". In: *WIREs Computational Molecular Science* 1.5 (2011), pp. 826–843. DOI: https://doi.org/10.1002/wcms.31.

[84] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. "Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method". In: *Phys. Rev. Lett.* 100 (2 Jan. 2008), p. 020603. DOI: 10.1103/PhysRevLett.100.020603.

[85] Giovanni Bussi and Alessandro Laio. "Using metadynamics to explore complex free-energy landscapes". In: *Nature Reviews Physics* 2.4 (Apr. 2020), pp. 200–212. ISSN: 2522-5820. DOI: 10.1038/s42254-020-0153-0.

[86] W. Kabsch and C. Sander. "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features". In: *Biopolymers* 22 (1983), pp. 2577–2637. DOI: 10.1002/bip.360221211.

[87] Gareth A. Tribello et al. "PLUMED2: New Feathers for an Old Bird". In: *Comput. Phys. Commun.* 185 (2014), pp. 604–613. DOI: 10.1016/j.cpc.2013.09.018.

[88] Giovanni Bussi, Davide Donadio, and Michele Parrinello. "Canonical Sampling through Velocity Rescaling". In: *J. Chem. Phys.* 126 (2007), p. 014101. DOI: 10.1063/1.2408420.

[89] M. Parrinello and A. Rahman. "Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method". In: *J. Appl. Phys. (Melville, NY, U. S.)* 52 (1981), pp. 7182–7190. DOI: 10.1063/1.328693.

[90] Miao Zhang et al. "Structural Basis for Calmodulin as a Dynamic Calcium Sensor". In: *Structure* 20.5 (May 2012), pp. 911–923. ISSN: 0969-2126. DOI: 10.1016/j.str.2012.03.019.

[91] https://github.com/ORBallesteros/TESIS/chapter2/cam.....COMPLETAR. Accessed: September of 2022.

[92] https://github.com/ORBallesteros/TESIS/chapter2/cam.....COMPLETAR. Accessed: September of 2022.

[93] https://github.com/ORBallesteros/TESIS/chapter2/cam.....COMPLETAR. Accessed: September of 2022.

[94] https://github.com/ORBallesteros/TESIS/chapter2/cam.....COMPLETAR. Accessed: September of 2022.

[95] Jun Zhang et al. "Mechanisms of ribosome stalling by SecM at multiple elongation steps". In: *eLife* 4 (2015). Ed. by Sjors HW Scheres, e09684. ISSN: 2050-084X. URL: https://doi.org/10.7554/eLife.09684.

[96] Schrödinger, LLC. "The PyMOL Molecular Graphics System, Version 1.8". 2015.

[97] Sunhwan Jo et al. "CHARMM-GUI: A web-based graphical user interface for CHARMM". In: *Journal of Computational Chemistry* 29.11 (2008), pp. 1859–1865. DOI: https://doi.org/10.1002/jcc.20945.

[98] Marija Liutkute, Ekaterina Samatova, and Marina V. Rodnina. "Cotranslational Folding of Proteins on the Ribosome". In: *Biomolecules* 10.1 (2020). ISSN: 2218-273X. DOI: 10.3390/biom10010097.

[99] https://github.com/ponnhide/pyCircos. Accessed: October of 2022.

[100] Po Wei Kang et al. "Calmodulin acts as a state-dependent switch to control a cardiac potassium channel opening". In: *Science Advances* 6.50 (2020), eabd6798. DOI: 10.1126/sciadv.abd6798.

[101] Sunhwan Jo, Taehoon Kim, and Wonpil Im. "Automated Builder and Database of Protein/Membrane Complexes for Molecular Dynamics Simulations". In: *PLOS ONE* 2.9 (Sept. 2007), pp. 1–9. DOI: 10.1371/journal.pone.0000880.

[102] Jumin Lee et al. "CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field". In: *Journal of Chemical Theory and Computation* 12.1 (Jan. 2016), pp. 405–413. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.5b00935.

[103] Jeffery B. Klauda et al. "Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types". In: *The Journal of Physical Chemistry B* 114.23 (June 2010), pp. 7830–7843. ISSN: 1520-6106. DOI: 10.1021/jp101759q.

[104] Seonghoon Kim et al. "CHARMM-GUI ligand reader and modeler for CHARMM force field generation of small molecules". In: *Journal of Computational Chemistry* 38.21 (2017), pp. 1879–1886. DOI: https://doi.org/10.1002/jcc.24829.

[105] Woody Sherman et al. "Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects". In: *Journal of Medicinal Chemistry* 49.2 (Jan. 2006), pp. 534–553. ISSN: 0022-2623. DOI: 10.1021/jm050540c.

[106] Schrödinger Release 2022-3: LigPrep, Schrödinger, LLC, New York, NY, 2021.

[107] G. Madhavi Sastry et al. "Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments". In: *Journal of Computer-Aided Molecular Design* 27.3 (2013), pp. 221–234. ISSN: 1573-4951. DOI: 10.1007/s10822-013-9644-8.

[108] Richard A. Friesner et al. "Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes". In: *Journal of Medicinal Chemistry* 49.21 (2006), pp. 6177–6196. ISSN: 0022-2623. DOI: 10.1021/jm051256o.

[109] https://github.com/ORBallesteros/TESIS/chapter2/cam.....COMPLETAR. Accessed: September of 2022.

[110] Csaba Böde et al. "Network analysis of protein dynamics". In: *FEBS Letters* 581.15 (2007), pp. 2776–2782. DOI: https://doi.org/10.1016/j.febslet.2007.05.021.

[111] Jeffrey R. Wagner et al. "Emerging Computational Methods for the Rational Discovery of Allosteric Drugs". eng. In: *Chemical reviews* 116.11 (June 2016). 27074285[pmid], pp. 6370–6390. ISSN: 1520-6890. DOI: 10.1021/acs.chemrev.5b00631.

[112] L. Di Paola et al. "Protein Contact Networks: An Emerging Paradigm in Chemistry". In: *Chemical Reviews* 113.3 (Mar. 2013), pp. 1598–1613. ISSN: 0009-2665. DOI: 10.1021/cr3002356.

[113] Mahdi Ghorbani, Bernard R. Brooks, and Jeffery B. Klauda. "Exploring dynamics and network analysis of spike glycoprotein of SARS-COV-2". eng. In: *Biophysical journal* 120.14 (July 2021). 33705760[pmid], pp. 2902–2913. ISSN: 1542-0086. DOI: 10.1016/j.bpj.2021.02.047.

[114] D. J. Klein. "Centrality measure in graphs". In: *Journal of Mathematical Chemistry* 47.4 (May 2010), pp. 1209–1223. ISSN: 1572-8897. DOI: 10.1007/s10910-009-9635-0.

[115] Andre A. S. T. Ribeiro and Vanessa Ortiz. "Determination of Signaling Pathways in Proteins through Network Theory: Importance of the Topology". In: *Journal of Chemical Theory and Computation* 10.4 (Apr. 2014), pp. 1762–1769. ISSN: 1549-9618. DOI: 10.1021/ct400977r.

[116] Andre A. S. T. Ribeiro and Vanessa Ortiz. "Energy Propagation and Network Energetic Coupling in Proteins". In: *The Journal of Physical Chemistry B* 119.5 (Feb. 2015), pp. 1835–1846. ISSN: 1520-6106. DOI: 10.1021/jp509906m.

[117] Wesley M. Botello-Smith and Yun Luo. "Robust Determination of Protein Allosteric Signaling Pathways". In: *Journal of Chemical Theory and Computation* 15.4 (Apr. 2019), pp. 2116–2126. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.8b01197.

[118] Ulrik Brandes and Daniel Fleischer. "Centrality Measures Based on Current Flow". In: *STACS 2005*. Ed. by Volker Diekert and Bruno Durand. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 533–544. ISBN: 978-3-540-31856-9.

[119] M.E. J. Newman. "A measure of betweenness centrality based on random walks". In: *Social Networks* 27.1 (2005), pp. 39–54. ISSN: 0378-8733. DOI: https://doi.org/10.1016/j.socnet.2004.11.009.

[120] Adam T. Van Wart et al. "Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis". In: *Journal of Chemical Theory and Computation* 10.2 (Feb. 2014), pp. 511–517. ISSN: 1549-9618. DOI: 10.1021/ct4008603.

[121] Annie M. Westerlund et al. "Network analysis reveals how lipids and other cofactors influence membrane protein allostery". In: *The Journal of Chemical Physics* 153.14 (2020), p. 141103. DOI: 10.1063/5.0020974.

[122] https://github.com/delemottelab/allosteric-pathways.

[123] Enrico Bozzo and Massimo Franceschet. "Resistance distance, closeness, and betweenness". In: *Social Networks* 35.3 (2013), pp. 460–469. ISSN: 0378-8733. DOI: https://doi.org/10.1016/j.socnet.2013.05.003.

[124] https://github.com/Kortemme-Lab/flex_ddG_tutorial. Accessed: September of 2022.